GENOMIC ANALYSIS OF MICRO-INVERSIONS BASED ON HIGH-THROUGHPUT SEQUENCING

A Dissertation
Presented to
The Academic Faculty

by

Li Qu

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
Department of Biomedical Engineering

Georgia Institute of Technology, Emory University and Peking University
May 2020

GENOMIC ANALYSIS OF MICRO-INVERSIONS BASED ON HIGH-THROUGHPUT SEQUENCING

Approved by:

Dr. Huaiqiu Zhu, Advisor
School of Biomedical Engineering
Peking University

Dr. Antony Chen
School of Biomedical Engineering
Peking University

Dr. May D. Wang, Co-advisor
School of Biomedical Engineering
Georgia Institute of Technology and Emory University

Dr. Minghua Deng
School of Mathematical Sciences
Peking University

Dr. Zuhong Lu
School of Biomedical Engineering
Peking University

Date Approved: [December 10, 2019]

To my family.

ACKNOWLEDGEMENTS

First of all, I would like to express my sincere gratitude to my main advisor Dr. Huaiqiu Zhu. Thank for his care and guidance in the past five years. I still remember that Dr. Zhu often worked late into the night to revise my paper. Every time I felt confused and anxious, Dr. Zhu taught me that I should have a serious research, a hard work and a steadfast attitude. From work to life, Dr. Zhu is an example for me to learn and a beacon on my way forward. As a student in the joint program of PKU/GT/Emory, I spent one year in the second campus Georgia Tech. I would like to thank my co-advisor Dr. May Dongmei Wang at GT for her valuable help in both research and life in the US.

Secondly, I would like to thank all my lab mates for their care and help. Academically, we got together to discuss problems and revise articles. Longshu Yang discussed with me once a week and guided the way forward. Luotong Wang, Feifei He, Yilun Han and Yiying Wang participated in my project and paid a lot of time and effort. Mo Li, Peihong Wang, Chunhui Wang, and Zhencheng Fang read my dissertation carefully and put forward suggestions for revision. I would also like to thank other colleagues for their help, especially Binbin Lai, Zhe Wang, Xiaoqing Jiang, Xin Li, Congmin Xu, Jie Tan, Zhongjie Xie, Man Zhou at PKU; and Li Tong, Ying Sha, Ryan Hoffman, Janani Venugopalan at GT. Finally, I would like to thank to my family and friends. Thanks to my parents for their tolerance, consideration and care. Thanks to my fiance for his silent support and love. I can't imagine how I can get through the difficult time and get the degree without their support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

Genomic structural variations (SVs) are generally defined to include insertions, deletions, duplications, translocations, copy number variations and inversions. As with other types of variations, inversions are of great significance for studying disease susceptibility, population diversity and human evolution. For the last few years, inversions have drawn increasing attention with the large amount of data produced by high-throughput sequencing. In this dissertation, we defined micro-inversions (MIs) as inversions with the length shorter than 100 bp and larger than 10 bp. Until now, there is still a lack of systematic analysis of MIs due to the following reason. The unmapped reads are usually completely discarded by previous SV detection tools and these unmapped reads may include SVs, mainly consisted of small-scale MIs, which could cause the reads to fail to map to the reference genome. Fortunately, the MI detection tool (MID) which used unmapped reads to detect MIs, was developed by our lab in recent years and made the MI analysis available. In this dissertation, we made a comprehensive systems biology analysis of MIs on both healthy genomes at the population level, and cancer genomes based on the MIs detected from high-throughput sequencing data. Specifically, we have accomplished the following two aspects of work:

(1) Based on the healthy individual genomes from the 1000 Genome Project (1KGP), we detected abundant MIs from 1,937 samples in 26 populations all over the world, built a landscape of MIs on non-disease individuals, and made a comparative analysis of MIs in non-human primate genomes from the University of California Santa Cruz (UCSC) Genome

Browser Database. Specifically, we discovered 6,968 MIs in human individuals with MID and 24,476 MIs in non-human primates including chimpanzee, gorilla, orangutan, gibbon, baboon, rhesus monkey, and squirrel monkey with searchUMI tool. The MI results in human genomes showed that MIs were rarely located in exon regions and the gene density might affect the MI distribution among chromosomes. Among the five super-populations, African had the most MIs and East Asian had the least, which was consistent with previous research on single nucleotide polymorphism (SNP). Furthermore, the average MI number among the five super-populations were in linear relationship with the descending order: Africa > America > Europe > South Asia > East Asia, and this descending order also coincided with the "Out of Africa" hypothesis, which assumed that humans originated in Africa and migrated to other continents later. Besides, Africans had the most MIs in common with non-human primates, which also supported "Out of Africa" hypothesis. The results of phylogenetic tree and PCA not only met our expectation but also reflected a regional pattern among the 26 populations suggesting that ethnic groups that live geographically closest to one another have a relatively small MI genetic distance. In addition, the cluster of MIs in the human populations also coincided with human migration history and ancestral lineage. Thus, we proposed that MIs were potential evolutionary markers for investigating population dynamics. In general, we made a comprehensive analysis of MIs in human genomes and our results revealed the diversity of MIs in human populations and showed that they were related to evolution, environmental adaptation, and health. These MI results may further support for the analysis of human genome diversity and the construction of human evolutionary process.

(2) Base on the cancer genome data from sequence read archive (SRA) database, we further detected and analyzed the MIs of 451 samples in six cancers, including esophageal cancer, bladder cancer, hepatocellular carcinoma, lung cancer, prostate cancer, and pancreatic cancer. We also used the 1,937 healthy individuals from 1KGP as control samples. We first analyzed the distribution of MIs among chromosomes in genomes of six cancers. The results showed that there were both similarities and differences in MI distributions among chromosomes in different cancers. We also found that the MI number in cancers was much higher than that in healthy samples. Besides, prostate cancer had the most MIs and hepatocellular carcinoma had the least. Moreover, we analyzed the genes with frequent MIs in six cancers. It showed that the genes in which MIs frequently appeared in different cancers were specific, and many of these genes were closely related to cancers. In addition, we compared the MIs we detected with the SNPs reported previously and found that 132 SNPs overlapped with MIs. In summary, our analysis of MIs in six cancers showed that the number of MIs in different cancers, as well as chromosome and gene preference, were different. The divergent MIs among six cancers may provide help for personalized diagnosis and therapy of the six cancers in the future.

In conclusion, based on the high-throughput sequencing data, we focused on studying the small-scale micro-inversions (MIs), which have been ignored for a long time, and made a comprehensive bioinformatics analysis of large amounts of MIs in human genomes. The analysis of MIs in healthy individual genomes may improve our understanding of human

genetic diversity and evolution. At the same time, through the comparative genomics analysis of MIs in different cancers, we hope to provide further understanding for precision medicine and revealing the disease mechanism

CHAPTER 1 INTRODUCTION

This chapter intends to introduce the motivation and background of this dissertation, which is mainly about micro-inversions (MIs) analysis in human genomes. As an important type of genome structural variations (SVs), MIs play important roles in disease susceptibility, population diversity and human evolution. Herein, we first give an introduction to SVs and MIs, and also describe the current research status about SVs and MIs. Since all the data we used for analysis in this dissertation are the whole genome sequencing (WGS) short reads based on high-throughput sequencing, we then give an introduce to the characteristics of the WGS short reads produced by high-throughput sequencing from 1000 Genome Project (1KGP) and Sequence Read Archive (SRA). As we used MID and searchUMI as the detection tools for identifying MIs, we finally give a brief description of the challenges of MI detection and introduce these two MI detection tools.

1.1 Introduction to structural variations and micro-inversions

1.1.1 The definition and research status of SVs

Structural variations (SVs) are one of the most important mechanisms that account for genetic diversity [1]. In addition, amounts of studies have shown that SVs account for the most important mechanism of evolution [2-6]. Generally, SVs include deletions, duplications, translocations, copy number variations and inversions as shown in Figure 1.1[7].

SVs generally include balanced variations and unbalanced variations[8]. The standard of classifying balanced variations and unbalanced variations is to see whether the sequence length changes after the variation. The length of sequence with balanced variations does not change after variation such as inversion and translocations, while the sequence length of unbalanced variations increases or decreases after variation such as insertions, deletions, duplications and CNVs.

Specifically, the insertion means that the individual genome contains a sequence that does not exist in the reference gnome; the deletion means that the individual genome missed a sequence on the reference genome; the copy number variation (CNV) or the duplication means that the individual genome contains multiple copies of a sequence on the reference genome; the inversion means that the individual genome contains a sequence that is the reverse complementary sequence on the reference genome. The specific details of the SVs including insertions, deletions, translocations, duplications, copy number variations (CNVs) and inversions are shown in Figure 1.1.

Statistical data show that SVs are quite heterogeneous in size, ranging from a few base pairs to several mega bases. The high-throughput sequencing provides sequencing reads of 100 bp which are much shorter compared with the reads of 700 to 1000 bp in Sanger sequencing. In fact, as one of the most important genome variations, SVs account for 1.2% of all the genome variations while the most studied single nucleotide polymorphisms (SNPs) only account for 0.1% [9]. In addition, the occurrence of the same or different types of SVs

simultaneously may cause very complex genome rearrangements[10]. SVs have been proven to be related with many diseases such as neurocognitive disorder, infantile autism, schizophrenia, colitis and cancer [11-14]. Numerous studies have shown that inversions as one of the important SVs, play significant roles in cancer susceptibility and human evolution[15].

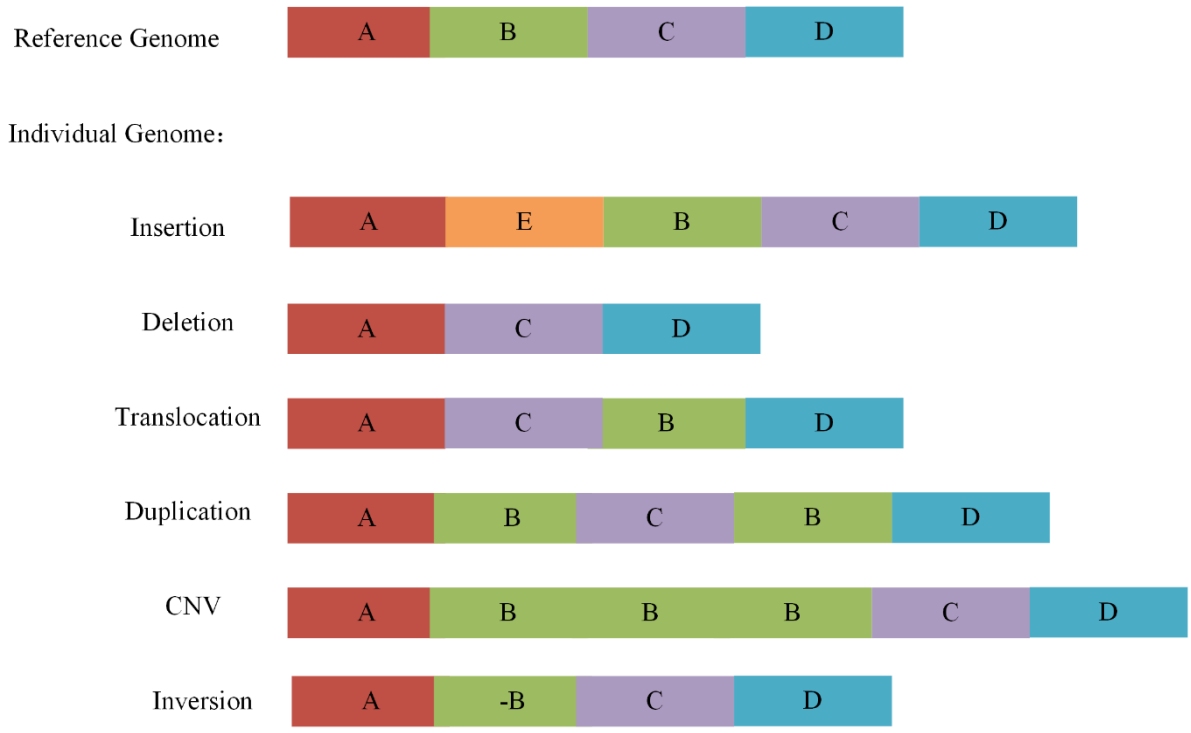

Figure 1.1 – Different types of SVs including insertion, deletion, translocation, duplication, copy number variation (CNV) and inversion. A, B, C, and D represent four consecutive DNA segments. E represents a DNA segment that is not consecutive to A, B, C, and D. "-B" in the schematic diagram of Inversion indicates a reverse of the original sequence segment B.

In the past few years, the studies of SVs have made a lot of progress with the advent of high-throughput sequencing[16]. Meanwhile, high-throughput sequencing also brought new problems and challenges. Indeed, high-throughput sequencing has provided plenty of data for metagenomics, protein binding regulation, gene annotation and genomic variations[17]. The

high speed and low cost of high-throughput sequencing have made huge amounts of individual genome sequencing data available.

By now, a few world-wide genome sequencing projects have been finished including 1KGP[18] and sequence read archive (SRA) databases[19]. These worldwide individual genome sequencing projects provide the possibility to better analyze the variation of the human genome structures. Moreover, these databases including 1KGP and SRA database are also beneficial to further understanding of human disease and precision medicine.

The study of gene variation started from 1936, while the research on human genome variation did not appear until 1977[20]. With the development of high-throughput sequencing technologies during these years, the analysis of SVs has been promoted fast and steadily. Additionally, high-throughput sequencing has provided plenty of data for studying metagenomics, protein binding regulation, gene function annotation, and multiple diseases including cancers [21-23].

Over the past decades, there have been growing interest in studying SVs of different scales[24,25]. Zhang *et al.* studied genome variations with different length size including SNPs (1 bp), fine-scale SVs (50 bp-5 kbp), intermediate-scale SVs (5 kbp-50 kbp), large-scale SVs (50 kbp-5 Mbp), and chromosomal SVs (>5 Mbp)[26]. Different scale of SVs have their own characteristics and affect the gene function in various aspects. Also, with the advent of high-throughput sequencing, a number of SVs databases began to appear, which summarized the analysis of SVs of current research institutions and scholars [18,19].

The effect of SVs on human genomes was first observed in healthy human individuals[27]. However, SVs have been proven to be related with various diseases including autism, schizophrenia, regional enteritis and cancers [11-14]. Specifically, Brandler *et al.* have proven that SVs are closely related to patients with autism[28]. Besides, some studies have showed that familial Parkinson's disease is caused by SVs in special genes, though most Parkinson's patients are sporadic[29,30]. In addition, Lakich *et al.* found that an inversion in VIII gene could cause hemophilia A and this inversion existed in about 43% of type A hemophilia patients[31].

SVs could generally account for the genetic diversity among different populations and affect gene expression in various ways. Some SVs could affect gene expression quantity by inserting or deleting some sequence copies such as insertions, deletions, and tandem duplications, while the other small-scale insertions, deletions or inversions may alter parts of gene sequences, thus cause damage of exons, alternative splicing and gene fusions[32].

As for the SVs in non-coding regions, they may influence regulatory elements and further affect gene function or signal pathways, Although SVs exist in many genes, it is reported that genome rearrangements especially the deletions are likely to lead to purifying selection or negative selection. Purifying selection or negative selection is the selective removal of alleles that are deleterious during evolution process[33]. Researches show that the genes enriched with SVs involved in many biological activities such as immune response, drug metabolism, signal transduction and sensory perception[34,35], which indicate that the study of SVs is of great significance.

The SV databases provide supports for the study of SVs and deeper understanding of the population genomes. However, these databases have the problem of incomplete and inconsistent SV information due to the fact that most SVs are obtained through different tools and experiments. Since there is currently no general guideline or tool for evaluating these SV sites, these SV-site information may not be absolutely reliable to a large extent.

Furthermore, the studies about the analysis of the common or overlapped SVs among different genomes are lacked. This lack of SV studies indicates that people's understanding of genomic SVs is still limited. Undoubtedly, future researches are needed to yield a more accurate and reliable database of such SV analysis among different genomes. To achieve this goal and to enrich disease-related genomic structural variation, future researches need to focus on how to construct a credible, accurate, genome-level overview of the human genome SVs.

Among all the various types of SVs, insertion and deletion are the most studied types, and the researches of deletion and insertion are also the most abundant in human genome sequences including in the cancer genomes[36-39]. In recent years, analysis regarding to the effects of copy number variations (CNVs) on genomes have been intensively studied[40,41]. However, as a balanced variation of SVs, inversions have not been studied a lot due to the detection difficulty. Thus, inversion occurrences and the effect of inversions are difficult to observe clinically unless they fall on gene regions or regulatory regions on the genome. Therefore, the comprehensive analysis of inversions is more urgent than other types of SVs.

1.1.2 Definition and the significance of micro-inversions (MIs)

An inversion is defined as a sequence reversal on the chromosome segment. Generally, SVs occur in different scales from a few base pairs to several mega bases[42]. In this dissertation, we analyzed micro-inversions (MIs) shorter than 100 bp and larger than 10 bp, which is defined in a previous work of our lab[43]. Inversions are an important type of genome variations, which are conductive to study human genomics and cancer genomes[44]. However, little is known about the fine-scale inversions, MIs, due to the limitation of tools of MIs until the MI detection tool, MID, was developed in a previous work by our lab [43].

Figure 1.2 shows the schematic diagram of an MI detected with the human assemble hg19 as the reference genome. In this dissertation. This MI occurred in gene ANKRD36 on Chromosome 2 from 97,826,002 to 97,826,027. So far, numerous studies have focused on the inversion diversity in human [45,46]. Since three decades ago, easily detectable macro-inversion polymorphisms in human have been verified by experiments and were implicated in discovery of human evolutionary history [47-48]. Flores *et al.* (2007) located three chromosomal inversions in tissue of human through frequent repeated DNA fragments undergoing non-allelic homologous recombination [49].

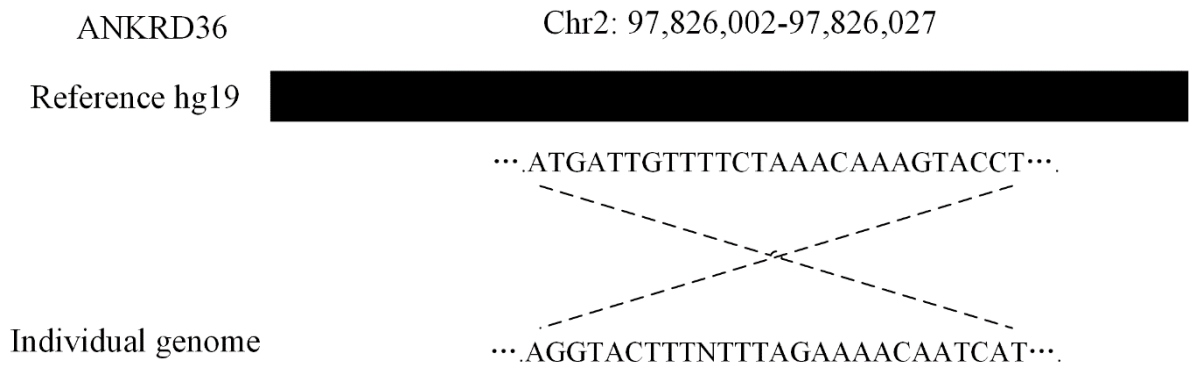

Figure 1.2 - One example of MI occurring in gene ANKRD36. This MI occurred on Chromosome 2 from 97,826,002 to 97,826,027.

With the development of sequencing techniques, the identification of inversion polymorphisms within species has been greatly promoted [50-52]. Cáceres et al. was able to precisely classify individuals in the population based on inversion status [53].The inversions located in 8p23.1, which are associated with autoimmune and cardiovascular disease [54], are indicated to be an evolutionarily marker in perceive of human phenotypic diversity [55]. Navarro et al. (2003) proved that large inversions might play a significant role in speeding up of the speciation of human and chimpanzee [56]. Osborne et al. demonstrated that a 1.5 million inversion commonly exist in families with Williams Beuren syndrome [57]. However, such studies on human inversions merely allowed the determination of variants of several kilo bases to mega bases due to the backwardness of experimental techniques [58-61].

In recent years, among all kinds of inversions, the small-scale inversions especially MIs with length of 10 to 100 bp have drawn increasing attention in the studies of SVs. Figure 1.3 displays an MI occurring in the 3 'UTR region of the gene PREPL and SLCA1 in the healthy

individual genomes from 1KGP reported by by He et al [42]. Both of genes PREPL and SLCA1 are reported as strongly related genes for Hypotonia-Cystinuria Syndrome (HCS) [62].

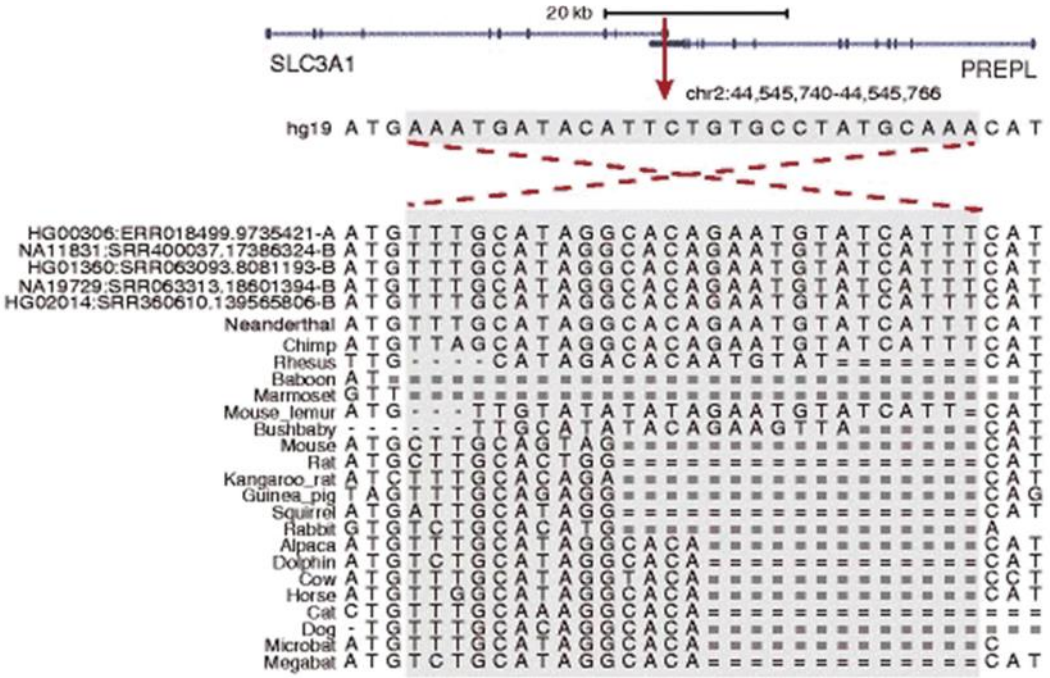

Figure 1.3 - An MI occurring in the 3' UTR region of gene SLCA1 and PREPL1[43]. This MI was found in 46 human individual genomes and some non-human primate genomes.

Besides, this MI existed in 46 individual genomes belonging to 14 different populations including JPT (Japanese in Tokyo, Japan), CEU (Utah Residents (CEPH) with Northern and Western Ancestry), FIN (Finnish in Finland), GBR (British in England and Scotland), TSI (Toscani in Italia), IBS (Iberian Population in Spain), CLM (Colombians from Medellin, Colombia), MXL (Mexican Ancestry from Los Angeles USA), PEL (Peruvians from Lima, Peru), PUR (Puerto Ricans from Puerto Rico), ACB (African Caribbeans in Barbados), ASW (Americans of African Ancestry in SW USA), LWK (Luhya in Webuye, Kenya) and YRI (Yoruba in Ibadan, Nigeria). Specifically, the 46 genomes in which this MI exist include two

individuals from East Asia, ten individuals from Europe, 12 individuals from America and 22 individuals from Africa. This MI existence in such many populations and continents implies the ubiquitous trend of MIs in human genomes.

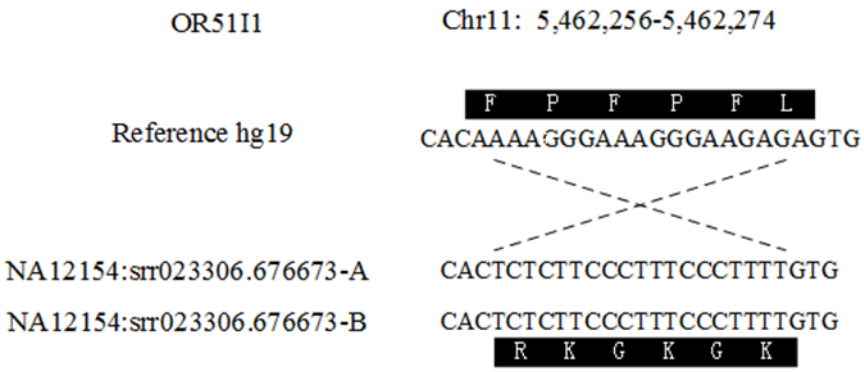

Figure 1.4 - An MI that alters six amino acid in gene *OR51I1* [43]. The six altered amino acids were located in the fourth transmembrane domain, which is less conserved than the extracellular regions.

As shown in Figure 1.3, at the same position with this MI on the chimpanzee genome and the Neanderthal genome, there was an almost identical sequences to the MIs found in the 46 human individual genomes. This result was likely to indicate that the reference human genome assemble hg19 has recently undergone this MIs compared with the corresponding sequence of the most recent common ancestor of human, chimpanzee, and the Neanderthal. However, the MI shown in Figure 1.3 was marked as multiple nucleotide variation by the existing genomic structural variation database (GeneBank: rs71416108) [63]. This result indicates that MIs may have been improperly reported as SNPs previously. We will further discuss the association of MIs and the SNPs proved previously in Chapter 3. It also implies that the understanding of MIs is not comprehensively enough as the result of too few researches on MIs. There is still a

huge room for improvement, which is also the significance of the study on MI analysis in this dissertation.

Another MI which altered six amino acids in the gene *OR51I1* was displayed in Figure 1.4[43]. The six altered amino acids were located in the fourth transmembrane domain, which is less conserved than the extracellular regions. Thus, the changes in these amino acids may have serious effects on health.

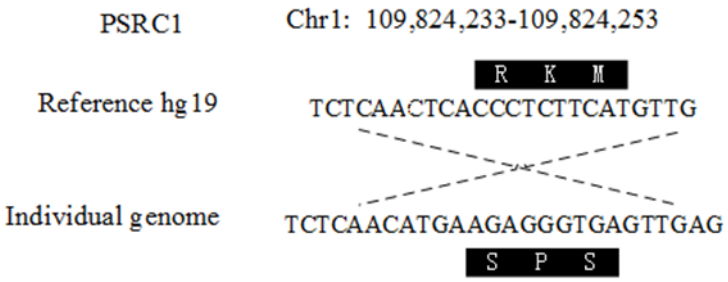

Figure 1.5 - An MI that alters three amino acid in gene PSRC1[43]. This MI interrupted the region boundary of gene PSRC1 and altered three amino acids in the coding region.

As is shown in Figure 1.5, an MI interrupted the region boundary of gene *PSRC1* and altered three amino acids in the coding region [43]. It should be noted that the *PSRC1* gene encodes a protein rich in proline, which is the regulatory target of p53. The MI occurrence in cancer- related genes may indicate the MI influence on cancer genome.

An MI occurring in the CDS region of gene *JMJD4* and the 5' UTR region of the gene *SNAP47* is shown in Figure 1.6[43]. Specifically, this MI alters five amino acids of *JMJD4* gene, which as a number of the Jmjc-domain-only families only contains JmjC domain. The JmiC domains play important roles in the process of demethylation and are closely related

with cancer diagnosis. Besides, the SVs in genes included in this family and the corresponding changed gene expression are reported to be associated with cancer regulation. Thus, this specific MI found in the CDS region of *JMJD4* gene may help further understanding the association of this gene and the diagnosis as well as treatment of related cancers.

Figure 1.6 - An MI that alters five amino acids. This MI occurrs in the CDS region of gene JMJD4 and the 5' UTR region of the gene SNAP47 [43].

Despite the clear evidence showing the direct relationship of MIs and cancers, the results above show that some MIs do overlap with CDS regions and regulatory regions, which may have an influence on human health. As for the MIs in intron regions, although the functions of these introns affected by MIs are not understood yet by now, we believe that these MIs will benefit studies of relation of variation and health in the future. Besides, the MIs shared only by a specific population especially those overlapped with gene regions, may increase the susceptibility of some disease. However, further sophisticated genetic and experimental data are necessary to confirm this point.

The high-throughput sequencing provides pair end sequencing reads of 100 bp which are much shorter compared with the 700 to 1000 bp in Sanger sequencing [17]. In this dissertation, we analyzed micro-inversions (MIs) with the length shorter than 100 bp and larger than 10 bp

[43], which is similar to the length of short reads produced in high-throughput sequencing. With the rapid development of high-throughput sequencing technology, plenty of short reads have raised new challenges for studying different-scale SVs especially MIs. For examples, MIs may cause sequence mismatches resulting in sequence assemble errors. Besides, inversions may have relation with gene evolution, genetic polymorphism and human health. Furthermore, compared with those large-scale inversions, MIs are more likely to occur [70]. Therefore, it is significant to explore the MIs and carry the comprehensive MI analysis.

1.1.3 The research status of MI studies

The result of numerous studies on human genome SVs indicates that, inversions as one of the most important types of SVs, have played a significant role in genetic diseases, cancer genomes, and human evolution[64,65]. However, inversions as a type of balanced variation, will not cause the DNA sequence length increase or decrease. Thus, it is hard to find out the direct impact of inversions unless they are located in genetic regions or regulatory regions[66]. In fact, very limited inversions are discovered on human genomes until 2005[67]. Later, the high-throughput sequencing and the high-quality human assembles have made the imbalanced variation inversion analysis possible for the reason that the reliable human assemble have provided more precise and stable standard for SV detection.

Unlike other SVs such as CNVs, deletions, and insertion, the analysis of inversions has been overlooked. Around 80% of deletions were reported to be included in the studies of SVs [7]. However, the analysis of inversions is still not adequate by now. As shown in Figure 1.7,

all archives of SVs from the databases of dbVar [68] and DGVa [69] exhibit an undeniable lack of recording inversions, particularly for micro-inversions (10 bp to 100 bp). This is not consistent with the inference proposed previously that smaller SVs are much more common than larger ones in the human genomes[70,71], which indicates that a majority of small inversions are possibly kept to be detected and analyzed.

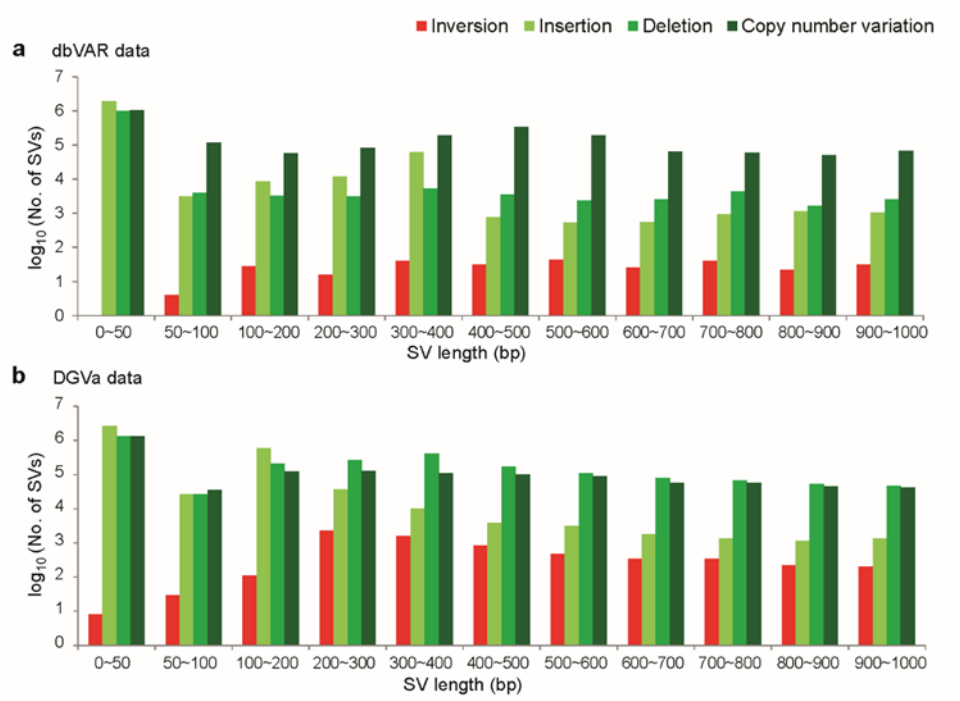

Figure 1.7 - All records on SVs from databases of dbVar and DGVa[72]. The records exhibit a lack of recording inversions, particularly for micro-inversions (10 bp to 100 bp).

During the process of SV detection and sequencing alignment, the sequencing short reads which cannot be mapped to the reference genomes are called unmapped reads. Generally, these unmapped short reads are completely discarded by previous SV detection tools and will not be processed later. In other words, the information in the unmapped reads is completely lost.

Actually, these unmapped sequencing reads may include SVs, mostly MIs, which could cause those reads to fail to map to the reference genomes. Due to the lack of alignment between these unmapped short reads and the reference genome, MIs have not been detected by other tools previously, let alone the MI analysis. We will further discuss the challenges of MI detection below.

In fact, people are not sure about how many inversions really exist definitively and how these inversions distribute. Besides, there is not a definite conclusion of the effect of MIs on human diseases [66]. Since the studies of MIs are very constrained, the mechanism of MI how MIs happen is not clear to researchers due to the following reasons. First, the high-throughput sequencing provides plenty of individual short reads, while the detection tools and analysis methods need to be changed to adapt such short reads. Second, MIs are balanced variations and most have not caused fatal damage to the genomes, which made the analysis of MIs extremely complex[73]. Last but not the least, there's not a reliable detection tool for identifying MIs until the MID [46] tool was developed by our lab.

Our analysis of MIs on actual genetic data including samples from 1KGP and SRA databases can fill the gap in the field of MI research. To better understand MIs, we included a comprehensive analysis of MIs with both healthy samples from 1KGP and cancer samples from SRA database. From the analysis of MIs in 1KGP and the comparison genome analysis of MIs between human and non-human primates, we acquired the distribution of MIs in both

population and ancestry scales. Besides, we obtained MI difference among different cancer genomes with data from SRA database.

1.2 The characteristics of the sequencing technologies

The current researches on SVs including our MI studies relies heavily on sequencing technologies. Furthermore, both the detection and analysis methods will change intensively with the rapid development of sequencing. In this dissertation, both the individual sequencing short reads used are whole genome sequencing (WGS) based on high-throughput sequencing. Besides, our understanding of MIs will alter with the development of sequencing technologies. The wide applications of high-throughput sequencing techniques have brought both problems and challenges to the studies of SVs including MIs. Based on this point, the sequencing technologies are of great importance. Thus, we would like to describe the characteristics of the high-throughput sequencing and WGS in this part.

1.2.1 The advantages and challenges brought by high-throughput sequencing

From the Sanger sequencing method, which appeared in the 1970s[74], to the rapid development of high-throughput sequencing technology in recent years, genome sequencing technology has been developed by leaps and bounds[75]. At the same time, the sequencing technology also greatly affects the study of SVs including MIs.

The working principle of Sanger sequencing is to segregate the sequence randomly into sequence fragments, and then introduce the sequence fragments into plasmid vector for

amplification to obtain enough sequence contents to be recognized by terminal staining. The first whole genome of human genomes was sequenced with Sanger sequencing technology[76]. The length of reads generated by Sanger sequencing are between 700 to 1,000 bp. This kind of sequencing is repeated several times to reach the required sequencing depth, which is the ratio of the total number of sequenced base pairs to the whole size of the sequenced genome[77]. However, the genome coverage by Sanger sequencing is usually not high, which makes the detection and analysis of SVs not convenient.

The mainstream sequencing technology currently is the high-throughput sequencing technology, of which the complete sequencing process includes building templates, sequencing imaging, data analysis and other steps[78]. In the sequencing process, complete DNA sequences are first broken up to screen for segments of a specific length (typically 100 bp). Then a sequence of tens to hundreds of base pairs is read from one end or both ends of each fragment. Although the length of each read is short, high-throughput sequencing technology can read a large number of such short sequences at the same time. This process makes the total length of all short sequences reach several to ten times of the length of the sample DNA, thus making it possible to obtain the high-coverage sample DNA sequence and detect SVs more effective.

In the high-throughput sequencing, the method of reading sequences from both ends of a fragmented DNA fragment is called pair-end sequencing. All the short reads used for analysis in this dissertation are pair-end sequencing. The piece of DNA being sequenced is a double-

stranded sequence, and the sequencing process consists of two directions starting at the 5' end of each strand of DNA. In the process of pair-end sequencing, the short reads are all paired, which could provide more information than single-end sequencing. Since the pair-end sequencing reads uses DNA fragments in two directions and there is an insert between the two ends, they are more widely used in detection of SVs especially the large-scale ones which can take good advantage of the insert size between the two ends.

The greatest advantage of the high-throughput sequencing technology is that it can obtain a large number of sequencing data in a short time. Therefore, there is a great improvement in sequencing time and material consumption in high-throughput sequencing compared with Sanger sequencing. The high-throughput sequencing platform depends on Implantable cyclic arrays to complete sequencing process. These arrays allow to produce at most millions of DNA sequencing reads at the same time. The plenty of high-throughput sequencing reads make the mass of SV researches possible.

The high-throughput sequencing has provided abundant data for metagenomics, protein binding regulation, gene functional annotation and gene structural variation[79]. Furthermore, the high speed and low cost of high-throughput sequencing has made massive individual sequencing data available. Accordingly, plenty of individual sequence databases have begun to emerge such as the 1000 genome project (1KGP) [18], and sequence read archive (SRA) databases [19]. These databases provide the possibility to better analyze SVs in human genomes and understand precision medicine in human cancer genomes.

However, the low coverage of some high-throughput sequencing data based on whole genome sequencing also increases the difficulty of MI detections. Most SV detection tools required the input file of sequencing short reads with sufficient coverage such as NA12878 released by 1KGP, which is widely used as a standard individual genome by SV detection evaluation and analysis. The SV detection results of NA12878 are also regarded as the gold standard of SVs (annotated data set) as a result of its high coverage sequencing. Despite of this, the sequencing coverage of the pair-end reads in the individual genome NA12878 is only 15x.

However, the actual situation is that except for a small number of high coverage individual genome samples, most of the genome reads in 1KGP are in low coverage of 2~4x. Few SVs detection methods can obtain satisfactory SV results with such a low-coverage data due to the limitations of the algorithm in using feature signals, such as the algorithm based on feature signals of read depth. But fortunately is, MI Detection Tool (MID) developed by our lab could take full advantage of these low-coverage individual genomes and still get a high sensitivity in detecting MIs. We will go into a deep discussion of the MID in the following parts.

The regularly used high-throughput sequencing platforms include Roche/454, Illumina, SOLID, Polonator, and HeliScope. These platforms differ in key sequencing processes and the corresponding sequencing reads [78]. However, the sequencing reads produced by these high-throughput sequencing platforms are commonly shorter than those in Sanger sequencing

platforms. Specifically, the sequencing reads produced by Illumina platform, used in this dissertation, are shorter than 100 bp which are called short reads. Hence, large amount of short reads produced by high-throughput sequencing, which have a high sequencing errors, need a new comprehensive analysis. Besides, the short length of the high-throughput sequencing brings new challenges in the detection of SVs, especially those extremely small scale SVs. Taking the personal genome sequencing data of 1KGP as an example, Illumina sequencing insert size in 1KGP is 100 to 600 bp, which greatly exceeds the length of MIs, which are generally 10 to 100 bp, so the pair-end mapped reads are not suitable for MIs detection in short reads under high-throughput sequencing.

1.2.2 The characteristics of whole genome sequencing (WGS)

Generally, WGS produce the complete DNA of the whole genome at a single time. WGS brings about sequencing of the genome chromosomal DNA along with DNA encompassed in the mitochondria[80]. WGS is often understood to be used to determine the human genome, but the scale and flexibility of the high-throughput sequencing is such that it can be used efficiently in any species, such as animal husbandry, plants, disease-related microorganisms and non-human primates[81-83].

Thus, whole genome sequencing is the most comprehensive approach to genome research. Genomic information could be used to identify genetic diseases, find mutations that drive cancer development and track disease outbreaks[84,85]. The rapidly falling cost of sequencing

and the increased ability to process large samples of data have made whole-genome sequencing the most powerful tool available to today's sequencers.

However, WGS has its own limitation. For example, WGS is time consuming and the cost of WGS is really high compared with whole exome sequencing (WXS)[86]. Besides, since the WGS needs to cover all the sites on the genome, the sequencing depth is usually not high.

At the same time, the low coverage of the high-throughput sequencing data also increases the difficulty of MI detections. Most SV detection tools required the input file of sequencing short reads with sufficient coverage, while the actual situation is that except for a small number of high coverage individual genome samples, most of the genome reads in 1KGP are in low coverage of 2~4x. Few SVs detection methods can obtain satisfactory SV results with such a low-coverage data due to the limitations of the algorithm in using feature signals, such as the algorithm based on feature signals of read depth.

Interestingly, the low-coverage sequencing genome data, which do not meet the requirement of many previous SV detection tools, are able to be used by MID. We will discuss this in the following parts. In this dissertation, all the data including the samples from 1KGP and SRA databases are whole genome sequencing. Although these data are in low coverage, the reliable detection of MIs with MID gives us a thorough landscape in both gene regions as well as the intergenic regions of human genomes.

1.3 The research progress of MI detection

1.3.1 The challenges of detecting MIs based on short reads

The advent of high-throughput sequencing has provided large amounts of short reads data to the SV analysis. However, the detection tools of the small-scale MIs are still constrained mainly due to the following reasons.

Most of the current existing SV detection tools rely on the first step of the sequence mapping. For example, the broadly studied database 1KGP used the sequence mapping tool Burrows-Wheeler alignment (BWA)[87] to first map the individual short reads to the human reference assemble to lay foundations for next steps such as SV detection. As the most widely used sequence alignment tool, BWA is designed based on the classic algorithm of Smith-Waterman. However, the Smith-Waterman algorithm could just simply deal with single-base mismatches and small-scale insertions or deletions, which are collectively referred as indels.

Nonetheless, the short reads in which MIs exist may fail to map to the reference genome. Generally, the sequencing short reads that are not able to map to the reference genome are called unmapped reads. During the mapping procedure, BWA attempted to reprocess these unmapped reads that could not be successfully matched in the first round by setting higher error rates for these reads. Nonetheless, they still could not handle the problem of mismatches caused by MIs. Thus, these unmapped short reads which include MIs are not be able to be parsed and are discarded finally.

Considering the reason described above, these unmapped short reads are usually completely discarded by alignment and mapping tools and are not processed later. Thus, the large amount of information that may be carried on the unmapped short reads is simply ignored by subsequent analysis tools and studies. Accounting that these unmapped short reads are not able to map to the reference genome, they do not contain important mapping information such as matching locations needed for the next SVs detection analysis. Therefore, most existing tools for analyzing SVs are not able to process these unmapped reads well. In fact, these unmapped reads may include SVs, possible mainly MIs, which cause these reads fail to map to the reference genome. In other words, if MIs could be found accurately, it is possible to find information carried in these reads that are ignored by other SV detection tools. However, For example, about 3% of the short reads in the individual genome NA19917 with the low-coverage sequencing from the publicly available database 1KGP are unmapped reads.

Besides, MIs are a kind of micro-scale structural variations, which are characterized by their short length and the complex detection from short reads. The length of short reads produced by the high-throughput sequencing makes it extremely difficult to determine MIs. The SVs detection tools such as CNVnator [88], which use read depth as the detection feature, have difficulties detecting MI events of such a small scale without being certain about breakpoints, thus have a low sensitivity. Besides, such detection tools are limited by the quality of the short read sequences and the type of SVs to identify. According to the types of SVs to

detect, those SV detection tools based on the feature signal of read depth are mostly used to detect deletions and CNVs and are not suitable for detecting MIs shorter than 100 bp.

Additionally, the SV detection tools based on the feature signal of read pairs such as DELLY[89] and those based on split reads such as Pindel[90] are not suitable for MI detection in short reads for the reason that the size of the SVs that such tools can detect depends on the insert size between paired-end short reads. Taking the personal genome sequencing data of 1KGP as an example, the Illumina sequencing insert size in 1KGP is 100 to 600bp, which greatly exceeds the length of MIs (generally 10 to 100 bp), so they are not suitable for MIs detection in short reads based on high-throughput sequencing. In addition, SV detection tools based on the feature signal of read pairs are very insensitive to SVs with a size as small as MIs, because they have difficulties distinguishing small disturbances from normal background changes with read pair spacing.

At the same time, the low coverage of the high-throughput sequencing data also increases the difficulty of MI detections. Most SV detection tools required the input file of sequencing short reads with sufficient coverage such as NA12878 released by 1KGP, which is widely used as a standard individual genome by SV detection evaluation and analysis. The SV detection results of NA12878 are also regarded as the gold standard of SVs (annotated data set) as a result of its high coverage sequencing. Despite of this, the sequencing coverage of the pair-end reads in the individual genome NA12878 is only 15x. However, the actual situation is that except for a small number of high coverage individual genome samples, most of the

genome reads in 1KGP are in low coverage of 2~4x. Few SVs detection methods can obtain satisfactory SV results with such a low-coverage data due to the limitations of the algorithm in using feature signals, such as the algorithm based on feature signals of read depth.

Studies have shown that population-level sequencing data can provide evidence of low-frequency SVs as well as more materials of relation between SVs and phenotypes [91]. Therefore, as for the amounts of low-coverage sequencing data from 1KGP we analyzed MIs at the both population-scale and ancestry-scale level by fusing the MIs in multiple individual genome samples in this dissertation. In this way, the integrated analysis data of MIs is more reliable compared those obtained at individual level.

Another difficulty of MI detection is the high false positive for the reason that there may be other multiple types of SVs around MIs in short reads, which may interfere with locating breakpoints and determining MIs events. Although MIs themselves as on kind of balanced variation, do not affect the length of the genome sequence, the other types of variations including SNVs, small indels, unbalanced SVs including insertions, deletions, duplications and CNVs, and other types of SVs including translocations may occur simultaneously in the MI regions and on the same short read.

In summary, MIs play significant roles in human genomes as discussed in the previous section, while MIs are still poorly understood by studies. The previous SV detection methods and tools have their own limitations on MI detection owing to such small scale and the complex situations around MIs on the genome. As for the MI detection in the assembled

genomes which are gathered by long contigs, the situation is much simpler. However, the mainly used sequencing reads in SV detection are the short reads which may not be simple to assemble. Despite the fact that the high-throughput sequencing has provided plenty of individual genome data, the personal genome database represented by 1KGP mostly provide low-coverage short reads which make MI detection challenging.

To solve the problems above, Micro-inversion Detector (MID) was developed by our lab in recent years. Besides, MID is the first known tool to perform MI detection analysis with unmapped short reads produced by high-throughput sequencing.

1.3.2 SearchUMI: detecting MIs from assembled genomes

Different with the lack of MI detection based on high-throughput sequencing, a few MI detection tools based on assembled sequences or the alignments against a reference genome have been proposed. Among of them, the MI detection tool SearchUMI [92], identifies MIs from the pair alignments of non-human primates against human reference genome hg19.

In this dissertation, for the seven non-human primate assembles, we applied the tool searchUMI, to detect MIs from the alignments between the human and seven non-human primate genomes. SearchUMI is able to identify inversions ranging from 5 to 125 bp, which is the approximate length of MIs. To make a better comparative analysis of MIs in human and non-human primate genomes, we select the MIs between 10 to 100 bp from the MI results of searchUMI.

SearchUMI identifies MIs from the pair alignments of non-human primate assemble genomes against hg19 in two steps: deciding the variation-rich genome regions and investigating the reverse locations from these variation-rich regions. The variation-rich regions are the regions where the gaps and mismatches are originally identified as a threesome of the closest gaps and mismatches that are nearer than supposed on the alignment. In case of that a gap or mismatch is discovered, the locations of the neighboring gaps and mismatches are locked. Finally, the reverse sequences are identified from these regions after discussion of each situation.

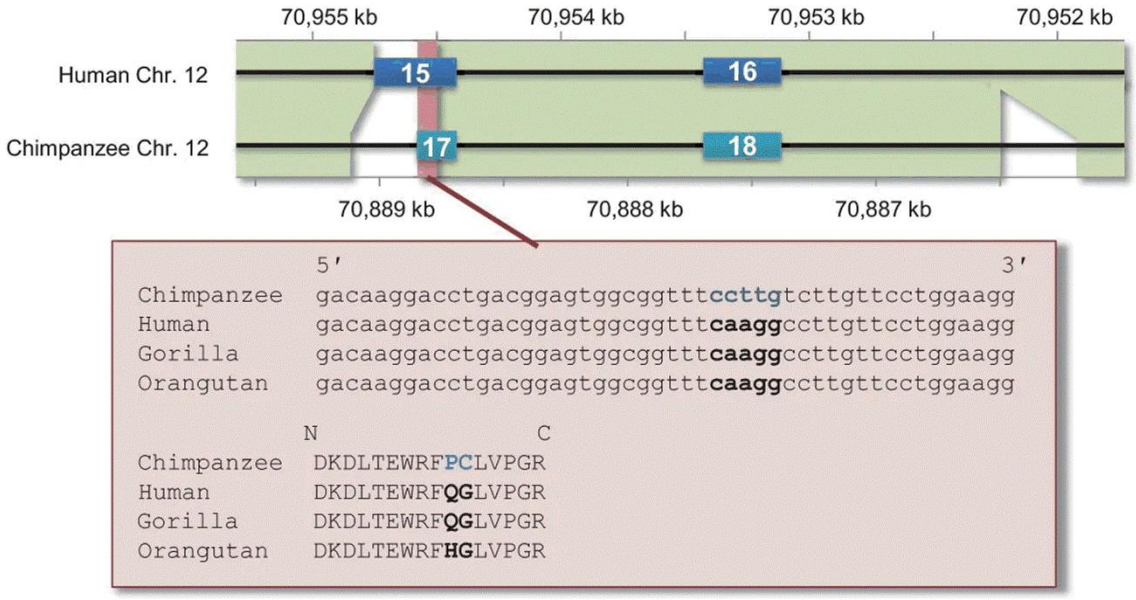

Figure 1.8 - An inversion of 4 bp found in gene PTPRB in alignment of chimpanzee assemble and human reference identified by searchUMI. The numbers represent the exon counts. The blue represents the MIs. The red box represents the alignment details of chimpanzee, human, gorilla, and orangutan[92].

It is reported that the PPV for searchUMI is more than 99.98 for detection of MIs in the sequences with balanced base distributions and 99.3% for the sequences with tendentious base

composition. The performance of searchUMI in the assemble alignment of non-human primates indicate that the MI results are credible. An inversion found in the pair alignment of chimpanzee and human altering two amino acids is shown in Figure 1.8. The inversion exist specifically in chimpanzee but not in gorilla or orangutan indicate MIs may have an effect on evolution in chimpanzee lineage. Thus the detection and analysis of MIs are of great significance.

Generally, SearchUMI is able to detect MIs from the alignments between the human reference genome assemble hg19 and seven non-human primate genomes including chimpanzee, gorilla, orangutan, gibbon, baboon, rhesus monkey, and squirrel monkey. Considering that the assemble of the non-human primates are more representative and comprehensive than any single primate individual, the comparative MI analysis between human and non-human primates are more convincing.

1.3.3 Micro-inversion detection (MID): detecting MIs from short reads

As described above, those unmapped short reads are usually completely discarded by alignment and mapping tools and are not processed later. Thus, the large amount of information that may be carried on the unmapped short reads is simply ignored by subsequent SV detection tools. In this dissertation, we used MID [43], which was proposed in a previous work of our lab in 2016, to detect MIs in both samples of healthy and cancer gnomes. Actually, these unmapped sequencing reads may include SVs, most likely mainly MIs, which could cause the reads not be mapped to the reference genome. Due to the lack of alignment of

unmapped reads and the reference genome, MIs have not been detected by other tools previously, let alone the MI analysis.

MIs are hard to detect and analyze due to their extremely short length compared with other kinds of SVs. In fact, the MI detection limitations of the previous SVs detection tools are mainly reflected in a few aspects. Firstly, most of the previous SV detection tools are insensitive to MIs, which is mainly because the length of MIs is too short. Besides, the MI detection process is also more complicated than that of other SVs. Secondly, many of the widely used personal genome sequencing data for SV detection are of low coverage, while most of the SV detection tools require high-coverage sequencing data. Finally, there may be other types of SVs around MIs on the same short read, which makes MI detection more difficult.

The previous work in our lab proposed MID (http://cqb.pku.edu.cn/ZhuLab/MID), which was the first tool to make use of unmapped reads that are ignored by other SV detection tools to detect MIs[43]. MID is able to match the fragments flexibly and deal with such problems as fragment overlaps and multiple breakpoints during MI detection process. In the MI detection process, MID initially remaps the unmapped reads onto the human reference assemble hg19 [116] by anchoring a sequence alignment. Then, MID gets all possible sequences matching fragment sets by a series of process including the fusion of k-mer, neighboring sub-sequence sets, removing the overlapped sequences and scoring these sequences. Among of these processes, scoring these sequences could effectively avoid false positives caused by

palindromic sequences or other mismatches. Finally, MID adopts dynamic algorithm to select the final optimum solution and outputs the reverse sequence information into the final output file. The pipeline of MID is shown in Figure 1.9.

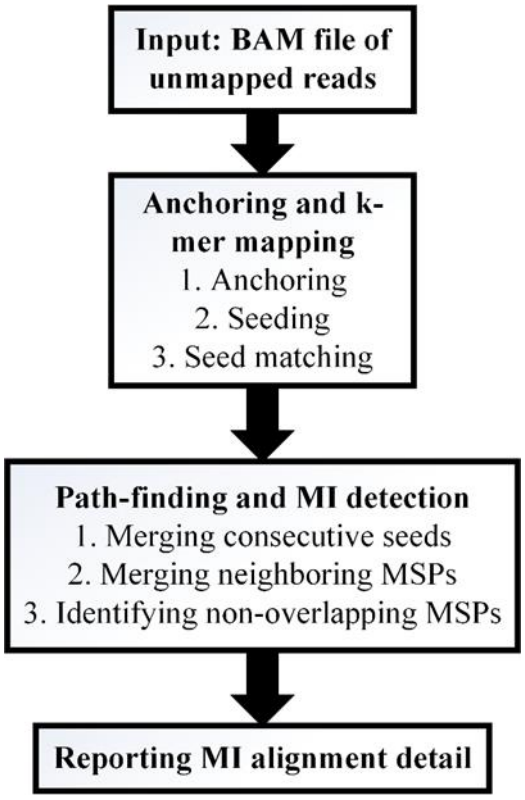

Figure 1.9 - The pipeline of MID [43]. The process of detecting MIs mainly includes inputting, anchoring and mapping, path-finding and MI detection, and reporting MIs.

MID generally outputs the MI information including the original short read name, chromosome, MI start coordinate, MI end coordinate, strand direction, MI length, and the specific sequence of MIs. The detailed information of the output file of MID is shown in Figure 1.10. In fact, there is no gold standard of human reference due to the genome polymorphism among different individuals. Usually, the human reference genome assembles such as hg19,

which are assembled by several individual genomes, are used as the alignment reference. Thus, MID also uses the hg 19 as the reference against the studied individual genomes to decide the exact MIs.

MID has a high accuracy and reliability when detecting MIs with the average sensitivity (SN) of 80.4% and Positive Predictive Value (PPV) of 90.2% in simulated sequencing data. Besides, MID performs well in both low-coverage and high-coverage sequencing reads. Moreover, MID also performs stably in the simulated data consisted of MIs as well as a mixture of other types of SVs. These results indicate that MIs detected by MID are reliable and believable. In general, MID is the first MI detection tool which used unmapped reads to detect MIs, and made the MI analysis on human genomes in this dissertation available.

```
@SRR063108.1193133 (HG01516.IBS)
s                hg19.chr5  179256626   100  +   180915260 AAG~AGATAGTTTTTTTATCTTTCCTCTCCTCCTAATG~TAG
s @SRR063108.1193133              0     72  +         100 AAG~AGA----------ATG~TAG
s @SRR063108.1193133             54     28  -         100 ----------TAGTTTTTTTATCTTTCCTCTCCTCCTA----------
```

Figure 1.10 - The format of the output file with detailed alignment information by MID [43]. This output file mainly includes the original short read name, chromosome, MI start coordinate, MI end coordinate, strand direction, MI length, and the specific sequence of MIs.

1.4 Introduction to the contents of this dissertation

SVs in human genomes have been studied extensively. However, little is known about the role of micro-inversions (MIs), generally defined as small (10 to 100 bp) inversions, playing in human evolution, diversity, and health. Portraying the pattern of MIs among

ethnically diverse populations is demanding for interpreting human evolution history, determining the appropriate design and giving us insight into genetic disease-related studies.

This dissertation is organized as follows:

Chapter 1 gives an overview of structural variations (SVs) and micro-inversions (MIs). With the advent of high-throughput sequencing, a large amount of high-throughput sequencing databases have emerged including 1KGP and SRA databases. These data provide possibility for the comprehensive analysis of SVs and the study of diseases. In this dissertation, MIs are defined as inversions shorter than 100 bp and larger than 10 bp. MIs are an important type of SVs, while the analysis of MIs is still lacked. Thus, this dissertation focusing on the MI analysis of both health samples and cancer samples is of great significance.

Chapter 2 describes the MI analysis in individuals from 1KGP. We explored the distribution of MIs in 26 human populations and seven non-human primate genomes, analyzed the phylogenetic structure of the 26 human populations based on MIs. We further investigated the functions of MIs located within genes associated with human health. The analyses of MIs in human genomes showed that MIs were rarely located in exon regions. The MI analysis of non-human primates and human populations was consistent with the "Out of Africa" hypothesis. The cluster of MIs in the human populations also coincided with human migration history and ancestral lineage. Thus, we proposed that MIs were potential evolutionary markers for investigating population dynamics. Our results revealed the diversity of MIs in human

populations and showed that they were related to evolution, environmental adaptation, and health.

Chapter 3 describes the MI distribution among six cancers from SRA database. To expand our knowledge of the roles of MIs in cancers, we analyzed the MIs of 451 samples from six cancers, including esophageal cancer, hepatocellular carcinoma, lung cancer, pancreatic cancer, prostate cancer, and bladder cancer. We also used the 1,937 samples from the 1KGP as the control samples. We further analyzed the distribution of MIs in the six cancers among 24 chromosomes. Besides the chromosome preferences, different cancers also have different preferences for various genes. We also calculated the average number of MIs per individual among each cancer. The MI preferences for divergent genes among six cancers may provide a guidance for the treatment and therapy on the six cancers.

Chapter 4 concludes this dissertation and prospects the future work.

CHAPTER 2 MICRO-INVERSIONS WITH CLUE TO POPULATION GENETICS ANALYSIS IN HUMAN GENOMES

MIs, as one of the most important SVs, may play important roles in human evolution, environment adaption, and health. This part describes the analysis of MIs in healthy human genomes on both the super-population and population scales. Our analysis of MIs with 1KGP data will improve our understanding of human genetic diversity and human evolution. The comparative analysis of MIs from the scale of populations, super-populations, and species will be the keystone of further implementation of human evolution theory.

2.1 Introduction

As a kind of SVs in genomes, inversions are defined as a chromosome reversal where a sequence segment ends upside down [52]. Usually, an inversion appears when a chromosome endures breakage and displacement within itself. Although it has been long known that inversions are associated with primate evolution [93], only recently they are found to play an important role in human evolution and diseases along with the wide applications of high-throughput sequencing techniques [94,95]. A number of studies have focused on inversions in the human genome [96] and for decades many detectable macro-inversion polymorphisms in humans have been verified by experiments and implicated them in human evolutionary history [97,98]. For example, inversions located in 8p23.1, which are associated with autoimmune and cardiovascular disease [99,100], have been designated as evolutionary markers of human

phenotypic diversity. Besides, a common 900-kb inversion polymorphism at 17q21.31 associated with Parkinson's disease suggesting multiple distributions of inversions among ethnic populations [95]. Indeed, with the developing of inversion detecting methods, inversions have been studied as one of the most important mechanisms accounting for genetic diversity.

However, the studies mentioned above were mainly limited to detecting the large-scale inversions, usually >100 kb. Recently several studies began to put more efforts using small-scale inversions with size much shorter than 10 kb into exploring such as phylogenetic problems. These studies found an influence of small inversions on forming unusual flanking sequences in human and chimpanzee genomes [101] and developed tools to detect pico-in-place inversions [102]. Nevertheless, their results have large discrepancies because of their non-uniform definitions based on the size of small inversions [43]. Moreover, current studies concentrate mainly on the small inversion differentiation of human (*Homo sapiens*) and chimpanzee (*Pan troglodytes*) genomes [101], while excluding the out-group primate genomes such as gorillas (*Gorrilla*), orangutans (*Pongo pygmaeus*), gibbons (*Hylobates sp*.), baboons (*Papio anubis*), and rhesus monkeys (*Macaca mulatta*) *et al*.

Of all inversions of small size, MIs, a type of extremely small inversions (generally 10 to 100 bp) found remarkably in human genomes, are uncertain with their function for the research community. However, statistical analysis implied that MIs, usually like other rare genomic changes, may serve as excellent phylogenetic markers because of their rare occurrence and low homoplasy [103]. Furthermore, many studies have examined the function

of MIs among multiple non-primate species such as yeast[104], sticklebacks[105], grasshoppers[106], drosophila[107], ducks[108], chicken[109], avian[110], and mice[111]. As for primates with larger genomes, comparative studies of the influence of chromosomal inversions occurring within or between chromosomes in the human and chimpanzee genomes can be traced back to the last century. Nevertheless, such studies on human inversions are limited because of the limitations of the detection techniques [112]. With the advent of high-throughput sequencing, 1KGP has provided a large number of WGS reads of healthy individuals on a large scale of populations across the world[114-116]. However, the short reads generated from high-throughput sequencing in the 1KGP data are generally <100 bp, which length are too short, leading to both the detection and analysis of the MIs difficult. Unlike the well-studied large inversions, which are easily detected, MIs have not yet been studied because it is difficult to detect MIs shorter than read length, most of which are identified as unmapped reads and are thus completely discarded. The MID[43] method, developed in our previous study, relies on unmapped reads for detecting MIs and performed well. The algorithm of MID is designed based on a dynamic programming path-finding approach which can efficiently and reliably identify MIs from unmapped short next-generation sequencing reads. This subsequently facilitates the analysis of MIs across a great scale of populations based on high throughput sequencing data. Therefore, an increased understanding of the MI landscape across various human races will lead to comparative analyses among individuals, ultimately providing guidance to precision medicine taking individual difference into account.

Although efforts have been devoted to analysis of small inversions in non-human creatures, there are not yet comprehensive studies of MIs, which are <100 bp, on human diversity, evolution and diseases in a large number of human genomes. In this study, we set out to detect MIs and further investigate the roles of MIs in the diversity and evolution of 26 human populations and seven non-human populations. Overall, we explored the distribution of MIs in all 26 populations from the 1KGP, and detected 6,968 MIs within all 1,937 human samples and 24,476 MIs in seven non-human primate genomes. From the detected MIs, we analyzed the extent of diversity of MIs and built phylogenetic trees by MI counts from both the scale of population and species. These results indicated that MIs rarely occurred in or nearby a protein-coding gene, and only a few were common in both primates and human populations. They also show that Africans share the most common MIs with other non-human primates, a finding that may provide evidence of the "Out of Africa" hypothesis [113]. More importantly, the phylogenetic analysis demonstrates that MI is a sensitive evolutionary marker for categorizing all populations. The categories coincide with human migration history and ancestral lineage. The analysis also implies the function of some MIs located within disease-causing genes. Thus, it is concluded that MIs should merit our attention for the studies of human evolution and environmental adaptation.

2.2 Materials and methods

2.2.1 Dataset

This part mainly introduces the datasets we used for the MI analysis of human and non-human primate genomes from 1KGP and UCSC Genome Browser.

1KGP is a genome sequencing project that aims to provide comprehensive personalized sequencing data of human genomes and construct the map of human genome SVs. 1KGP sequenced thousands of individuals from different countries and nationalities around the world and got a massive amount of personalized genomes. In 2012, 1KGP committee first released both the read sequencing and SVs of 1,092 individuals around the world[114]. Three years later in 2015, 1KGP released the recent SV results of 2,054 samples[115].

The recent samples in 1KGP are from 26 populations around world. They are CDX (Chinese Dai in Xishuangbanna, China), CHB (Han Chinese in Beijing, China), CHS (Southern Han Chinese), JPT (Japanese in Tokyo, Japan), KHV (Kinh in Ho Chi Minh City, Vietnam), BEB (Bengali from Bangladesh), GIH (Gujarati Indian from Houston, Texas), ITU (Indian Telugu from the UK), PJL (Punjabi from Lahore, Pakistan), STU (Sri Lankan Tamil from the UK), CEU (Utah Residents (CEPH) with Northern and Western Ancestry), FIN (Finnish in Finland), GBR (British in England and Scotland), IBS (Iberian Population in Spain), TSI (Toscani in Italia), CLM (Colombians from Medellin, Colombia), MXL (Mexican Ancestry from Los Angeles USA), PEL (Peruvians from Lima, Peru), PUR (Puerto Ricans from Puerto Rico), ACB (African Caribbeans in Barbados), ASW (Americans of African Ancestry in SW USA), ESN (Esan in Nigeria), GWD (Gambian in Western Divisions in the

Gambia), LWK (Luhya in Webuye, Kenya), MSL (Mende in Sierra Leone), and YRI (Yoruba in Ibadan, Nigeria).

Indeed, 1KGP classified all the 26 populations according to their ancestry lineage into five super-populations including East Asia, South Asia, Africa, Europe, America. Specifically CDX, CHB, CHS, JPT, and KHV are classified into East Asian Ancestry (EAS); BEB, GIH, ITU, PJL, and STU are classified into South Asia Ancestry (SAS); CEU, FIN, GBR, IBS, and TSI are classified into European Ancestry (EUR); CLM, MXL, PEL, and PUR are classified into American Ancestry (AMR); ACB, ASW, ESN, GWD, LWK, MSL, and YRI are classified into African Ancestry (AFR). Among these samples, 1,937 samples which are found with MIs are included for analysis in this dissertation.

The SV polymorphism across all the 26 populations released by the 1KGP Consortium is shown in Figure 2.1. The size of each pie chart is proportional to the population polymorphism in the corresponding population. Each pie chart is divided into four parts: the private deep color represents SVs private to this population; the private light color represents SVs private to this continent; light grey color represents SVs shared across the continent where the population resides in; deep grey color represents SVs shared across all continents. Our results on MI analysis also show that some MIs are shared among the five super-populations and the other MIs are private to one specific super-population or private to a population. In general, each population has both private MIs and common MIs shared with other populations in the same super-population.

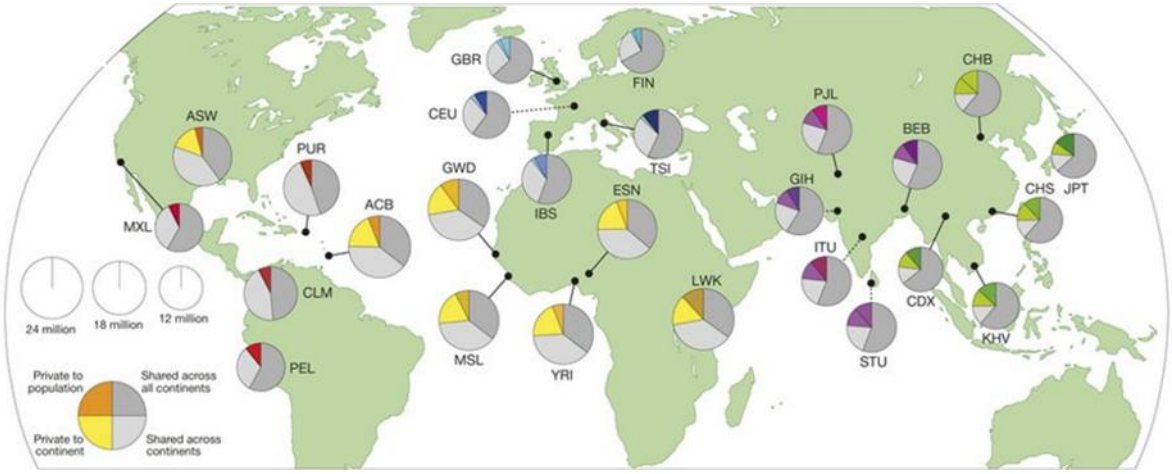

Figure 2.1 - SV polymorphism shown from 1KGP. The size of each pie chart is proportional to the population polymorphism in the corresponding population. Each pie chart is divided into four parts: the private deep color represents SVs private to this population; the private light color represents SVs private to this continent; the light grey color represents SVs shared across the continent that the population in; the deep grey color represents SVs shared across all continents[116].

In recent years, there have emerged many researches focusing on SVs using 1KGP data. Trevor *et al*, have used 1KGP data to explore the relationship between deleterious variations and disease risks [117]. Ting *et al* studied APOL gene variation and the haplotype diversity with 1KGP data [118]. The study by the 1KGP Consortium shows that abundant SVs are located in gene regulatory regions including untranslated regions (UTRs), promoters, enhancers, insulators and transcription factor binding sites (TFBS) [116].

The variant site number per genome of the 26 populations is displayed in Figure 2.2. It shows that the variation number of different populations varies a lot, which also coincides with our MI result. In this dissertation of MI analysis, we also find that the populations of African ancestry including ACB (African Caribbeans in Barbados), ASW (Americans of African Ancestry in SW USA), ESN (Esan in Nigeria), GWD (Gambian in Western Divisions in the

Gambia), LWK (Luhya in Webuye, Kenya), MSL (Mende in Sierra Leone), and YRI (Yoruba in Ibadan, Nigeria) have high MI occurrence frequency. This result also supports the "Out of Africa" hypothesis [119]. Moreover, the populations mixed with multiple ancestry lineages tend to have more MIs compared with those with single ancestries. The variation number from the 1KGP shown in Figure 2.2 displays that ASW which have American and African ancestry have very high variation frequency. Besides the variation number in the populations of American ancestry including CLM (Colombians from Medellin, Colombia), MXL (Mexican Ancestry from Los Angeles USA), PEL (Peruvians from Lima, Peru) and PUR (Puerto Ricans from Puerto Rico) rank only second after African populations. These could also be due to the African ancestry of these American populations.

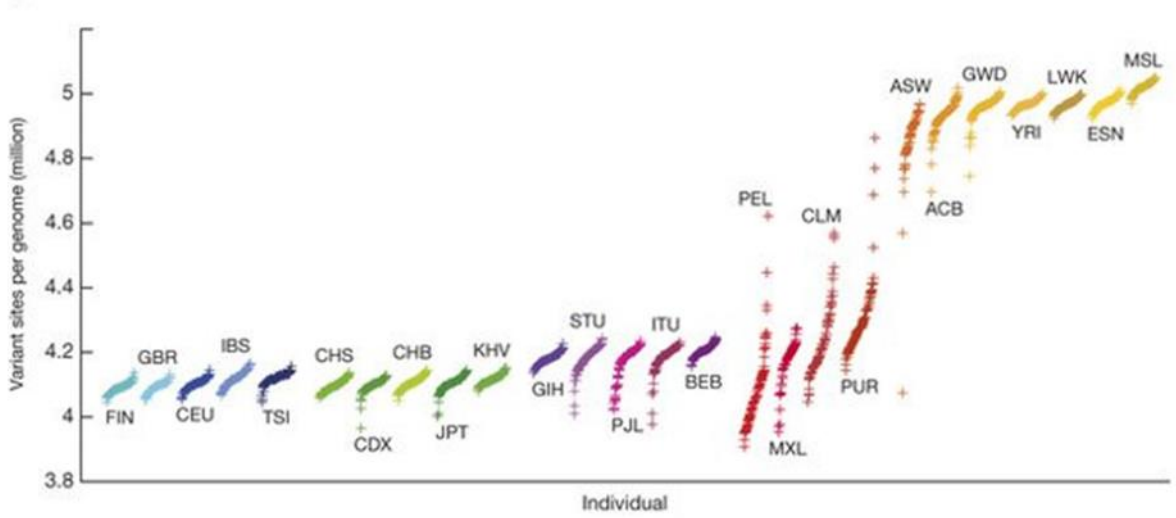

Figure 2.2 - The variant site number per genome of 26 populations from 1KGP[116].

To elucidate the features of MIs among human genomes, we collected WGS BAM files of the unmapped reads of 2,427 samples from the most recent version of the 1KGP [116]. 1KGP provided plenty of individual genome sequencing data based on high-throughput sequencing

and laid a solid foundation for SV studies. Moreover, the data of 1KGP has been widely used in the exploration and researches of SVs. However, there have not been studies about MIs with 1KGP data. Due to the limitations of current sequencing technology, sequencing time and sequencing cost, the coverage of most individual genome data is still very low. Among them, the average coverage of individual low-coverage genome sequencing data is 2x~4x. These low-coverage data have provided both chances and challenges for the analysis of MIs. Fortunately, MID could take good advantage of these low-coverage data and detect important MIs.

Then, MID was used to deal with all the low-coverage unmapped short reads using human assemble hg19 as the reference genome and gave a list of the detailed MI information. Of all the 2,427 samples, 490 did not contain MIs and were excluded. The included 1,937 samples, which covered all of the 26 populations, were categorized into five super-populations: East Asians [CDX (Chinese Dai in Xishuangbanna, China), CHB (Han Chinese in Beijing, China), CHS (Southern Han Chinese), JPT (Japanese in Tokyo, Japan), and KHV (Kinh in Ho Chi Minh City, Vietnam)], South Asians [BEB (Bengali from Bangladesh), GIH (Gujarati Indian from Houston, Texas), ITU (Indian Telugu from the UK), PJL (Punjabi from Lahore, Pakistan), and STU (Sri Lankan Tamil from the UK)], Europeans [CEU (Utah Residents (CEPH) with Northern and Western Ancestry), FIN (Finnish in Finland), GBR (British in England and Scotland), IBS (Iberian Population in Spain), and TSI (Toscani in Italia)], Americans [CLM (Colombians from Medellin, Colombia), MXL (Mexican Ancestry from Los Angeles USA),

PEL (Peruvians from Lima, Peru), and PUR (Puerto Ricans from Puerto Rico)], and Africans [ACB (African Caribbeans in Barbados), ASW (Americans of African Ancestry in SW USA), ESN (Esan in Nigeria), GWD (Gambian in Western Divisions in the Gambia), LWK (Luhya in Webuye, Kenya), MSL (Mende in Sierra Leone), and YRI (Yoruba in Ibadan, Nigeria)]. The sample number of each population used in this dissertation is displayed in Figure 2.3. Besides, the details of all samples from 1KGP are listed in Appendix A.

For further comparative analysis between humans and other primates, we also downloaded the human reference genome assembly hg19 [116] and the pairwise alignment data of seven non-human primates from the UCSC Genome Browser Database [120] (http://hgdownload.soe.ucsc.edu/downloads.html). The UCSC Genome Browser Database include a broad collection of vertebrate and model organism assemblies. On June 22, 2000, UCSC and the other members of the International Human Genome Project consortium completed the first working draft of the human genome assembly, forever ensuring free public access to the genome and the information it contains. A few weeks later, on July 7, 2000, the newly assembled genome was released on the web at http://genome.ucsc.edu , along with the initial prototype of a graphical viewing tool, the UCSC Genome Browser.

In this dissertation, we included assemble alignments from seven non-human primates including chimpanzee, gorilla, orangutan, gibbon, baboon, rhesus and squirrel monkey. Since the assembly of each primate is gathered from several individual genomes, the assembled alignment is more comprehensive than any individual primate genome. Considering that there

are many versions of each primates, we listed the assemble version for each primate here. We used the data version panTro4 for the human/chimpanzee alignment, gorGor3 for the human/gorilla alignment, ponAbe2 for the human/orangutan alignment, nomLeu1 for the human/gibbon alignment, papHam1 for the human/baboon alignment, rheMac3 for the human/rhesus alignment, and saiBol1 for the human/squirrel-monkey alignment from the UCSC genome browser (http://genome.ucsc.edu/). The detailed information of the seven non-human primate assembles is shown in Table 2.1. The columns of Primate, Synonyms, Coverage, Sequence length, Scaffolds, and Contig represent primate names of the alignment, the synonyms of the alignment version, sequencing coverage depth, the total sequence length of the alignment, the number of scaffolds used for the assemble alignment and the number of contigs respectively.

2.2.2 MI detection and annotation

To detect MIs from the 1KGP data, we applied the software MID[43] with default parameters and mapped all unmapped sequencing reads to the human assembly hg19. The choice of MID was prompted by our previous investigation that MID is capable of efficiently identifying MIs from unmapped short reads through inversion reading and reference genome mapping. MID could give the list of MIs and detailed information of each MI including chromosome, MI start coordinates on the chromosome, MI end coordinates on the chromosome, the length of MI and the exact corresponding MI sequence. Specifically, the

same MI occurring in multiple short reads of the same individual was counted as one during MID detection.

Table 2.1 - The seven non-human primate assemble alignments.

Primate	Synonyms	Coverage	Sequence length	Scaffolds	Contig
Chimpanzee	panTro4	6x	3,309,561,368	27,002	183,859
Gorilla	gorGor3	2.1x /35x	3,029,537,234	53,823	461,501
Orangutan	ponAbe2	6x	3,437,863,358	79,309	408,185
Gibbon	nomLeu1	5.6x	2,936,035,333	17,968	17,968
Baboon	papHam1	High coverage	/	/	/
Rhesus	rheMac3	50x	2,969,971,616	38,224	448,689
Squirrel monkey	saiBol1	80x	2,608,588,537	2,686	151,414

After detecting MIs, we annotated them with the gene via the GENCODE database[121], GENCODE is a genetic annotation database that identifies genetic characteristics through a series of computational analyses, manual annotations, and experimental results. GENCODE integrates a lot of gene information such as protein coding, long noncoding RNA (lncRNA), coding sequence (CDS) and is widely used in genome annotation. During the process of gene annotation, MIs were annotated as intergenic, genetic, exon, intron, CDS and UTR. In case that an MI was annotated as intergenic, it means this MI was in intergenic region and didn't overlap with any gene regions; if an MI was annotated as gene, it means this MI was overlapped with one gene on the hg19; if an MI was annotated as exon it means that this MI was overlapped with on exon of one gene; if an MI was annotated as intron, it means that this MI was in gene regions but didn't overlap with any exon of this gene; if an MI was annotated

as CDS, it means that this MI overlapped with coding sequence region of one exon; if an MI was annotated as UTR, it means that this MI overlapped with untranslated regions of one gene.

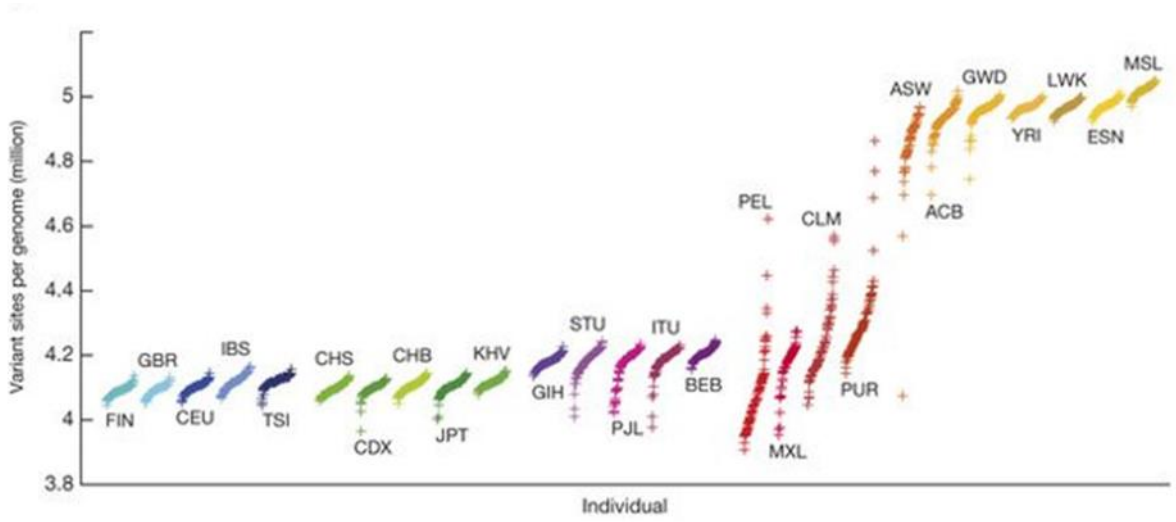

Figure 2.3 - The individual genome number of the 26 populations

The gene function and the correlated translated protein functions were annotated with GeneCards [122] well as Metascape [123]. GeneCards is a searchable, integrative database that provides comprehensive information on all human genes. It integrates gene information including gene function, the corresponding protein function and signal pathway functions. We annotated the private genes of super-populations that MIs were frequently located in with the protein function and further explored the gene function and specific health problems. Metascape is an online tool for retrieval of gene functions. We used Metascape as a gene function supplement to the annotations from GeneCards.

We detected the MIs of non-human primates by searchUMI tool, on aligned data because it is able to identify inversions ranging from 5 to 125 bp, which is the approximate length of

MIs. During detection, the parameter pd was set at 0.0137 for panTro4, 0.0175 for gorGor3, 0.034 for ponAbe2, 0.029 for nomLeu1, 0.066 for papHam1, 0.065 for rheMac3, and 0.123 for saiBol1. Furthermore, based on the definition of MIs in this dissertation, we removed inversions that were <10 bp.

Although the recurrence of specific MIs is rare, a series of MIs often converge to particular regions along the genome. To make a more comprehensible analysis of MIs in the following sections, by extending the boundaries of overlapping MIs, we defined MI regions (MIRs) as the union regions of overlapping MIs and MIRs need to meet the requirement that MIRs are at most 4 bp longer than the MIs they contain (shown in Figure 2.4), similar to the copy number variation regions (CNVRs) proposed by Yang *et al* [124]. According to our definition of MIR, it turns out that each MI is contained in one and the only unique MIR and each MIR may contain one or more MIs. Thus, the MIs contained in the same MIR, have almost the same start and end position on the chromosome.

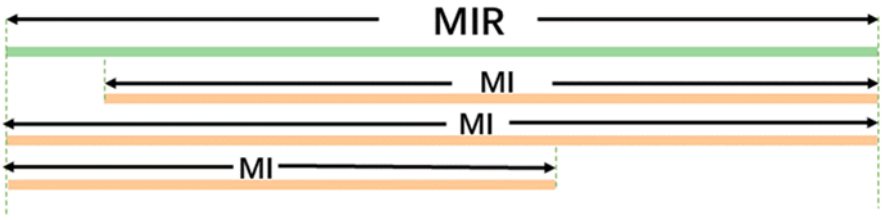

Figure 2.4 - Schematic showing how MIR refers to the region of a union of overlapping MIs (green bars, MIR; yellow bars, MIs in one individual).

2.2.3 MI diversity and population structures.

In order to analyze the diversity of MIs among all primates from an evolutionary and comparative viewpoint, we built matrices with rows denoting MIRs and columns denoting all 26 human populations. Each element in this matrix represented the sum of MIs included in the corresponding MIR within this population. In light of these matrices, we constructed a phylogenetic tree with the neighbor-joining algorithm via the R computer language. MIR sharing was analyzed by counting the number of shared MIRs among the five super-populations. The distances used in the neighbor-joining algorithm were Euclid's distances between every pair of column vectors, i.e., the distance between two populations was defined as the average of pairwise MIRs between two individuals from the two populations.

MIR sharing was analyzed by counting the number of shared MIRs among the five super-populations. The population structure of the 26 populations was visualized through PCA with R language to figure out if the distribution of MIs is on account of the geographical locations and migration history. In the PCA, we used the same matrix as that for the phylogenetic tree analysis. In addition, we regarded 2,140 MIRs as the principle components.

2.3 Results

Our previous work (He *et al.*, 2016)[43] has focused on the detection of MIs in the initial phases of 1KGP, which is however, encompassed by fewer individuals from only 19 populations and is largely limited to the statistical analysis of MI distribution, in spite of a few disease-related genes analysis. Herein we constructed a more integrated map of MIs including extra and new samples from 1KGP phase III data.

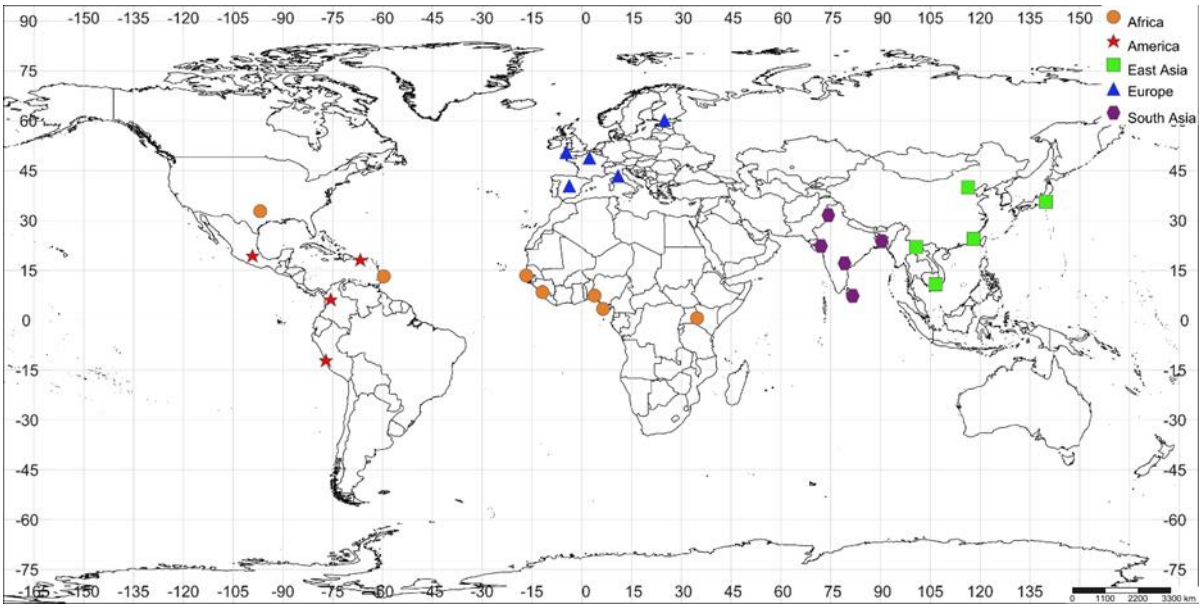

Figure 2.5 - Geographic locations of the 26 populations from 1KGP. Different color fills represent the five super-populations: Orange, Africa; red America; blue Europe; purple, South Asia; green, East Asia.

Besides, compared with our previous work, this study, made a more comprehensive analysis which emphasized the population diversity of MIs, quantified the MI genetic impact, and explored the important role that MIs play in human evolution. In this dissertation, we included 1,937 samples, which covered all of the 26 populations, categorized into five super-populations: East Asians (CDX, CHB, CHS, JPT, and KHV, South Asians (BEB, GIH, ITU, PJL, and STU), Europeans (CEU, FIN, GBR, IBS, and TSI), Americans (CLM, MXL, PEL, and PUR), and Africans (ACB, ASW, ESN, GWD, LWK, MSL, and YRI). 6,968 MIs were detected from the total 1,937 samples. Among 6,968 MIs, 1,149 are from East Asia, 1,387 from Europe, 2,146 from Africa, 850 from America, and 1,436 from South Asia. The geographic locations of the 26 populations from 1KGP is shown in Figure 2.5. Different color fills represent the five super-populations: orange, Africa; red America; blue Europe; purple,

South Asia; green, East Asia. The pie chart of the MIs in five super-populations is displayed in Figure 2.6.

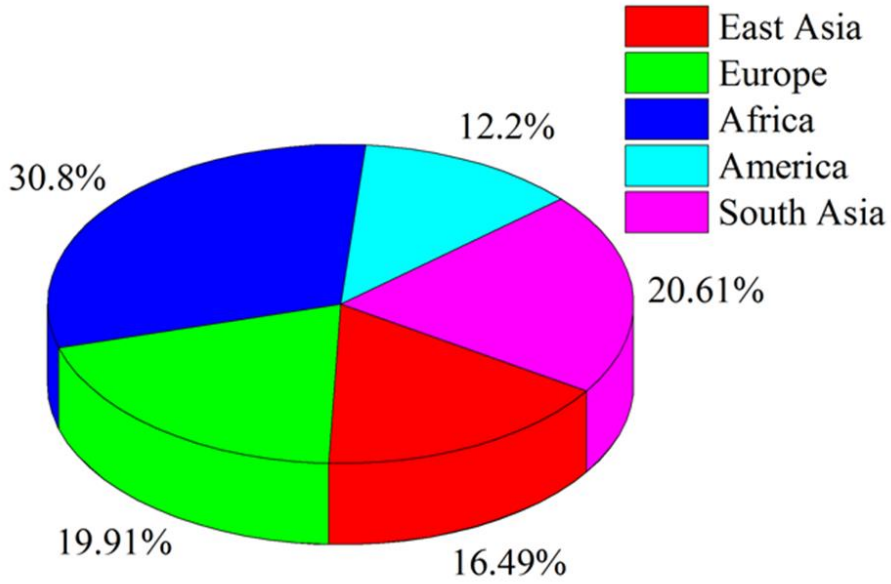

Figure 2.6 - Pie chart of MIs of MIs in fiver super-populations. The proportion of MIs in East Asia, Europe, Africa, America and South Asia are 16.49%, 19.91%, 30.8%, 12.2%, and 20.61% respectively.

2.3.1 Overview and distribution of MIs in 1KGP

With the MID method, we identified 6,968 MIs in 1,937 samples from the 26 human populations, and they were merged into 2,140 MIRs. The Table 2.2 list all the 6,968 MI information. The table is organized by populations and super-populations. The columns of the table from the left to right represent population or super-population abbreviations, population description, number of samples from 1KGP, MIR number, MI number, the number of MIRs supported by at least two MIs, the ratio of the number of MIs over the number of samples, which indicates average MIs per individual, and the ratio of multiple MIs supported MIRs

over the number of all MIRs. As shown in the table, the average MI number of each individual is 3.6. Besides, 977 of the 2,140 MIRs (45%) are supported by at least two MIs.

Table 2.2 - Overview of MIs detected in 1937 samples.

Pop	Population Description	Sam-num	MIR-num	MI-num	Mul-sup	MI-num/ Sam-num	Mul-sup /MIR-num
CHB	Han Chinese in Beijing, China	73	119	202	30	2.77	0.25
JPT	Japanese in Tokyo, Japan	90	209	416	61	4.62	0.29
CHS	Southern Han Chinese	80	116	183	27	2.29	0.23
CDX	Chinese Dai in Xishuangbanna, China	62	59	99	16	1.6	0.27
KHV	Kinh in Ho Chi Minh City, Vietnam	74	176	249	32	3.36	0.18
EAS	East Asia	379	483	1149	166	3.03	0.00
CEU	Utah Residents (CEPH) with Northern and Western Ancestry	80	300	387	37	4.84	0.12
TSI	Toscani in Italia	75	137	195	23	2.60	0.17
FIN	Finnish in Finland	77	237	283	18	3.68	0.08
GBR	British in England and Scotland	74	185	251	31	3.39	0.17
IBS	Iberian Population in Spain	70	199	271	31	3.87	0.16
EUR	Europe	376	877	1387	140	3.69	0.00
YRI	Yoruba in Ibadan, Nigeria	50	156	247	33	4.94	0.21
LWK	Luhya in Webuye, Kenya	83	161	278	44	3.35	0.27
GWD	Gambian in Western Divisions in the Gambia	102	177	384	65	3.76	0.37
MSL	Mende in Sierra Leone	81	168	316	29	3.90	0.17
ESN	Esan in Nigeria	87	167	347	51	3.99	0.31
ASW	Americans of African Ancestry in SW USA	58	149	293	50	5.05	0.34
ACB	African Caribbeans in Barbados	81	152	281	51	3.47	0.34
AFR	Africa	542	634	2146	323	3.96	0.00
MXL	Mexican Ancestry from Los Angeles USA	55	93	176	26	3.20	0.28
PUR	Puerto Ricans from Puerto Rico	39	107	220	46	5.64	0.43
CLM	Colombians from Medellin, Colombia	70	163	267	31	3.81	0.19

Table 2.2 (continued)

PEL	Peruvians from Lima, Peru	63	125	187	22	2.97	0.18
AMR	America	227	347	850	125	3.74	0.00
GIH	Gujarati Indian from Houston, Texas	83	108	193	29	2.33	0.27
PJL	Punjabi from Lahore, Pakistan	85	141	351	54	4.13	0.38
BEB	Bengali from Bangladesh	68	131	270	36	3.97	0.27
STU	Sri Lankan Tamil from the UK	91	139	334	51	3.67	0.37
ITU	Indian Telugu from the UK	86	119	288	53	3.35	0.45
SAS	South Asia	413	377	1436	223	3.48	0.00
Total	-	1,937	2,140	6,968	977	3.60	0.12

Pop, population name; **Sam-num**, number of samples for the population; **MIR-num**, number of MIRs in each population; **MI-num**, the number of MIs in each population; **Mul-sup**, the MIRs supported by at least two MIs; **MI-num/Sam-num**, the ratio of the number of MIs over the number of samples, which indicates average MIs per individual; **Mul-sup/MIR-num**, the ratio of multiple MIs supported MIRs over the number of all MIRs.

The result indicates that MIs are shared by different individuals, populations, and even super-populations. In general, Table 2.2 listed the real distribution of MIs among human genomes. In this dissertation, MIs are displayed in many levels including short reads, individual genomes, population-scale, and ancestry-scale. Furthermore, the MIRs supported by multiple short reads, individuals genomes, various populations will contribute to the concrete understanding of MIs.

Among the 6,968 MIs, there are five appearing in more than 200 individual genomes. The five high-frequency MIs may exist in many healthy people, which means that these MIs may also be considered to be added into the human reference assemble hg19. Since hg19 is still not complete, and new-found genome variations are usually added when updating hg19, the five high-frequency MIs could help human reference genome updating in the future as

presented in SV studies24. Of the 2,140 MIRs, 1,169 ones (54.6%) overlapped with gene regions, of which 1,063 (90.9%) overlapped with intron regions, while notably the rest 106 with exon regions for gene regions. Of the 106 MIRs in exon regions, 30 (28.3%) overlapped with Coding Sequence (CDS) regions, 39 (36.8%) with Untranslated Regions (UTR), and 37 (34.9%) with other functional regions including miRNA, Mt_rRNA *et al*.

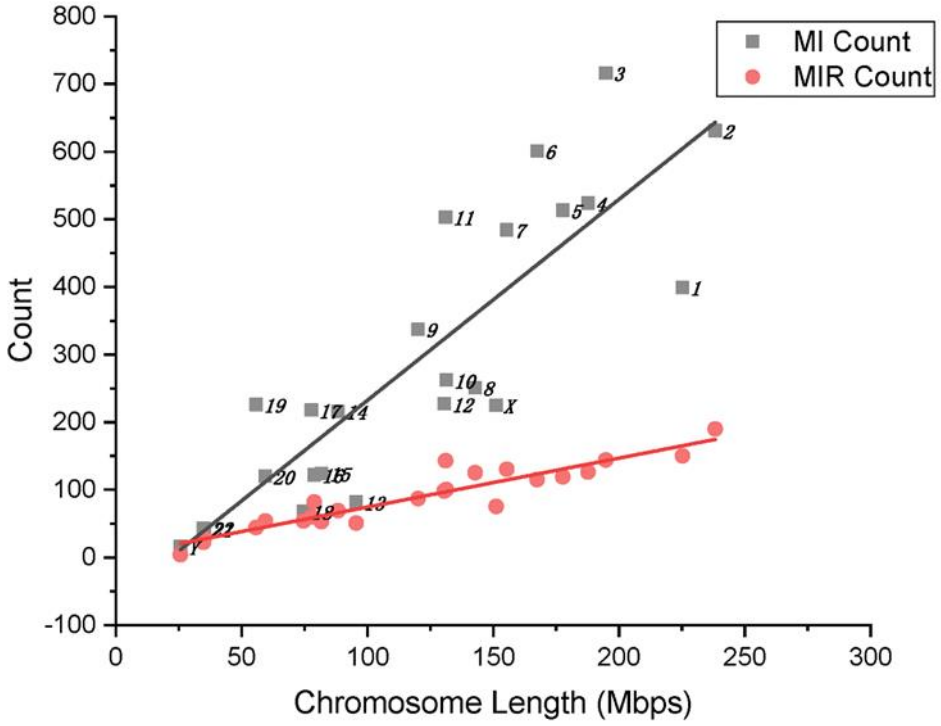

Figure 2.7 - Scatter plot of MIR count against chromosome length. The numbers from 1 to 22 and the characters X and Y represent the 24 chromosomes (22 autosomes and two sex chromosomes).

Indeed, the MIs in CDS regions could be important candidates for future studies, for the reason that compared with MIs in intergenic regions or intron regions, those MIs in the exon regions, especially in the CDS regions, are more likely to change the protein sequence directly and correspondingly cause phenotypical changes.

The scatter plot for MI and MIR count against the length of chromosomes is shown in Figure 2.7. Generally, the number of MIs and MIRs are positively correlated with the lengths of chromosomes (Figure 2.7), which is reasonable as longer chromosomes has more DNA bases, thus have more chances to make an error in DNA replicating.

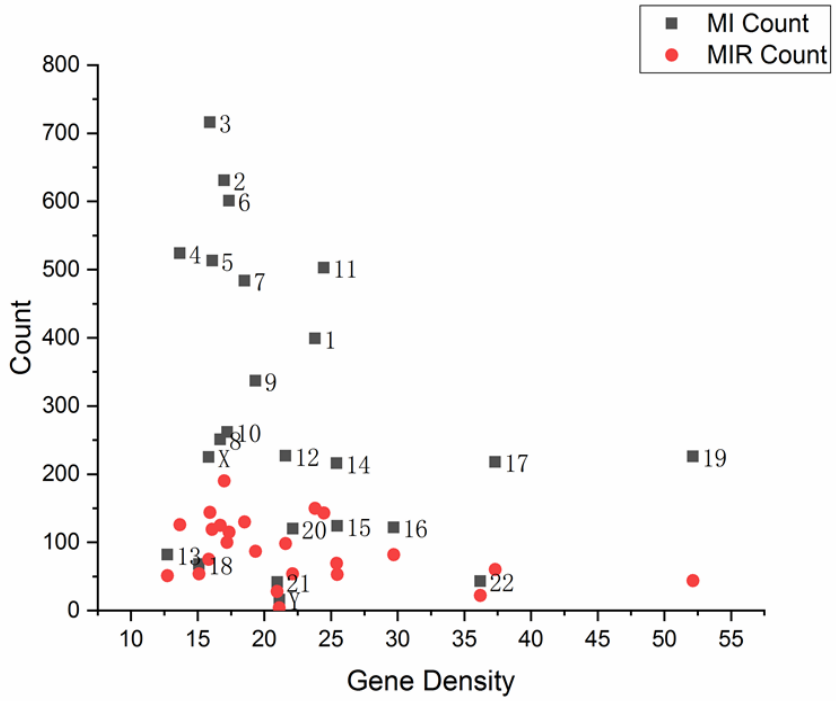

Figure 2.8 - Scatter plot of MI and MIR count against gene density.

Additionally, correlation between MI or MIR count and chromosome length was analyzed using Pearson's correlation analysis. Specifically, strong correlation was detected between MIR count and chromosome length, where the Pearson's correlation coefficient was r=0.935 (P<0.0001); the corresponding result for MI was 0.861 (P<0.0001). The scatter plot of MI and MIR count against the gene density was also described in Figure 2.8. Although we found no strong linear correlation between the number of MIs and gene density, we found a phenomenon that most scatters above the fitting line in the Figure 2.7 were located in the

chromosomes with high gene density. For example, we discovered that the chromosomes 19, 17, 16, and 11, of which the corresponding scatters were above the fitting line, have high gene density, as shown in Figure 2.8. This suggests that gene density may also affect the MI count of a chromosome, while this relationship is not strictly linear. We also displayed the distribution of MI and MIR event rate distribution across chromosomes in Figure 2.9.

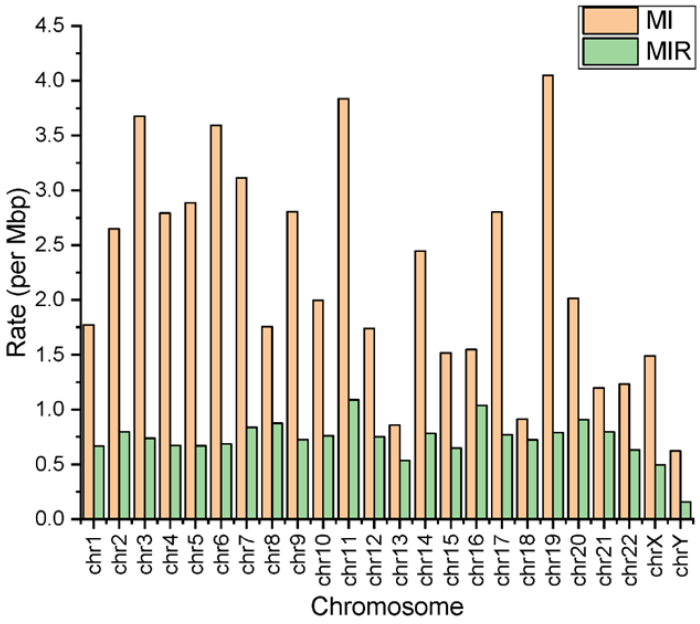

Figure 2.9 - The distribution of MI and MIR event rate distribution across chromosomes.

The Length distribution of 6,968 MIs and 2,140 MIRs is shown in Figure 2.10. As shown in Figure 2.10, the lengths of the MIs varied from 15 to 45 bp, but were concentrated within 18 to 30 bp. The specific positions of MIs among 24 chromosomes are shown in Figure 2.11. It should be noted that comparing the MIs in our study with the inversion-calling results from the phase III analysis of the 1KGP, shortest inversion detected by the 1KGP was 257 bp, which

was much larger than 100 bp. We also noted other researches about inversions and no MIs in the range of 10 to 100 bp were discussed previously. This suggests that there is no length overlapped between MIs in this dissertation and traditional inversions in the previous studies. Thus, our analysis on MIs (<100 bp), overlooked by all the studies including the 1KGP analysis, will increase our limited understanding of human SVs.

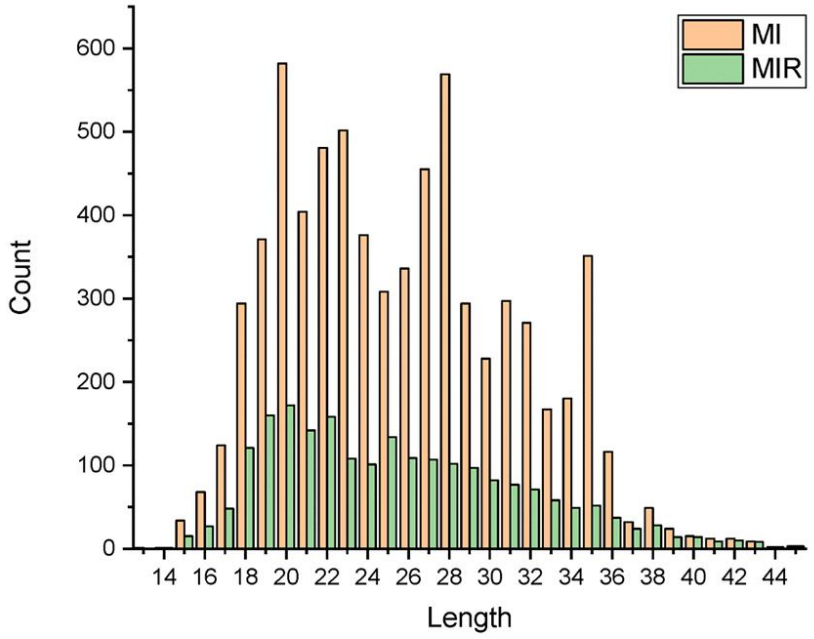

Figure 2.10 - Length distribution of 6,968 MIs and 2,140 MIRs. The length of MIs range from 13 to 45 bp and the length of MIRs range from 14 to 45 bp.

2.3.2 MI count per individual among 26 populations

SNPs and SVs in the 1KGP have been reported to display various allele frequencies and reveal genetic diversity among 26 populations [116]. However, MIs, as a kind of SVs, have never been studied to capture the genetic diversity among 26 populations. Thus, we performed the diversity of MI analysis in this part.

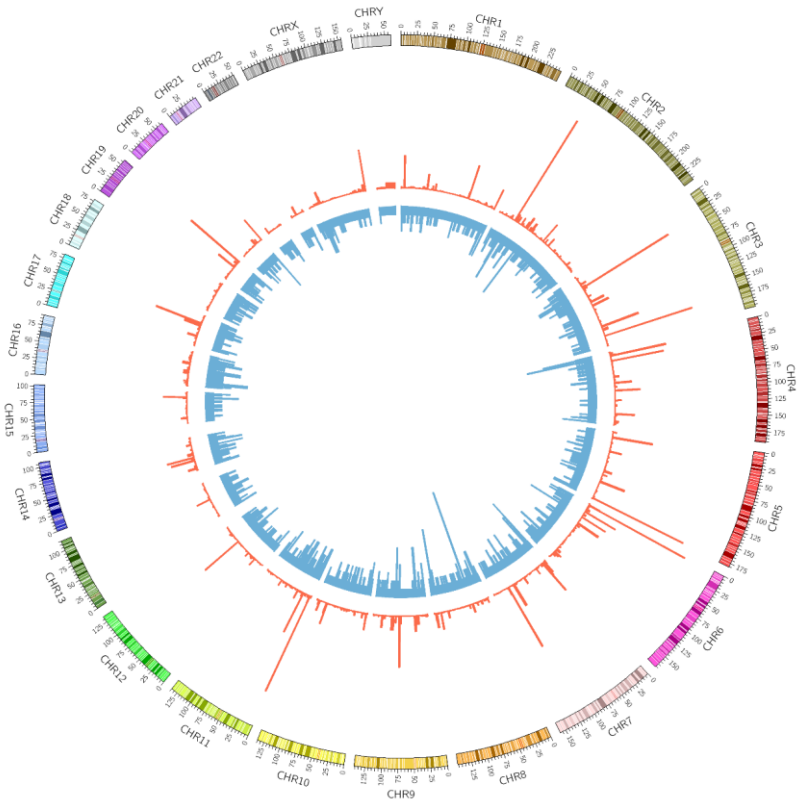

Figure 2.11 - Distribution of locations of MIs and MIRs on chromosomes. The outermost ring coordinates correspond to each chromosome. The number on the axis represents the coordinates of the bases on the chromosome in Mbp. The red and blue bars represent the MIs and MIRs.

In view of the fact that the sample number varies by population, diversity is better reflected by the average number of MIs per individual among super-populations and populations than the total MI number of each super-population and population. To address this point, we define parameter $\overline{C_p}$ to represent the average count of MIs per individual among super-populations or populations as follows:

$$\overline{C_p} = \frac{\sum_{i=1}^{N_p} C_i}{N_p} \quad (1)$$

where C_i means the count of MIs per individual, and N_p means the total number of individuals in a population or super-population.

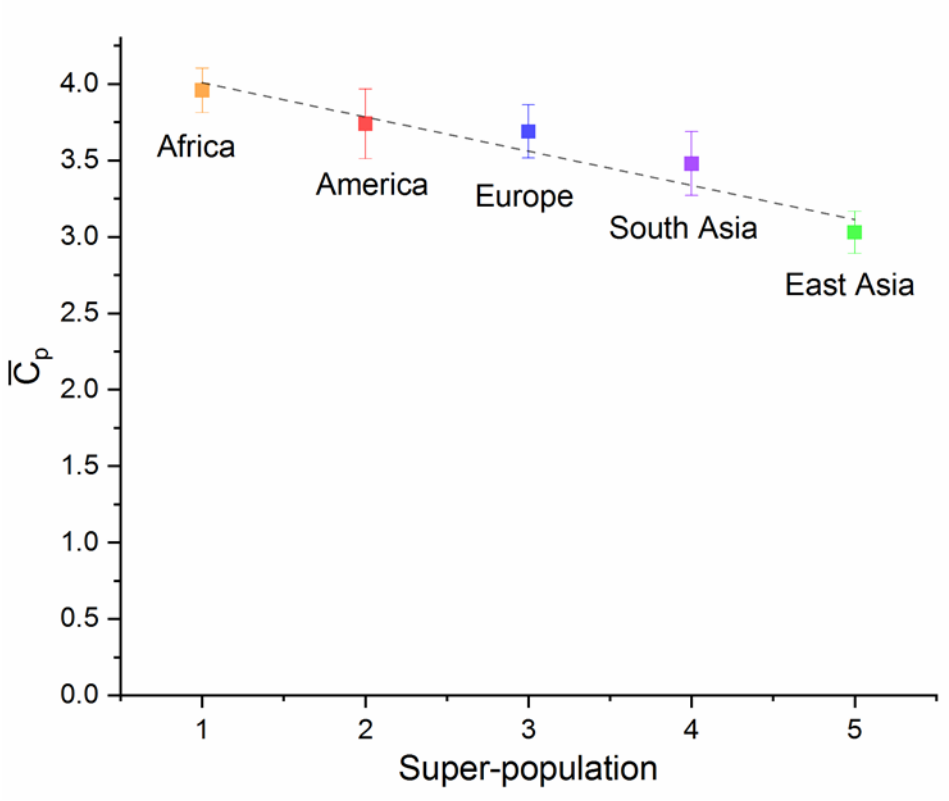

Figure 2.12 - Error bar plot of average count of MIs per individual among super-populations with fitted regression line. The data are presented as the mean ± SE. $\overline{C_p}$ is the average count of MIs per individual. Different color fills represent the five super-populations. Orange, Africa; red America; blue Europe; purple, South Asia; green, East Asia respectively. The average count ranged from 3.03 to 3.96.

From the aspect of the individual, we calculated an average of 3.6 MIs per individual. The error bar plot with the fitted regression line for the average number of MIs per individual among the five super-populations including Africa, America, East Asia, Europe and South Asia was shown in Figure 2.12. Indeed, we numbered the five super-populations, Africa, America, Europe, South Asia, and East Asia from one to five and performed Pearson's correlation analysis among the five super-populations. The Pearson's correlation coefficient was calculated, r=–0.967, P<0.01. This analysis showed the descending order of the average

count of MIs per individual among the five super-populations: Africa > America > Europe > South Asia > East Asia. This descending order supported the “Out of Africa” hypothesis, which believed that humans originated in Africa, then modern Africans migrated to the other continents. We will discuss this situation further in the part of Discussion.

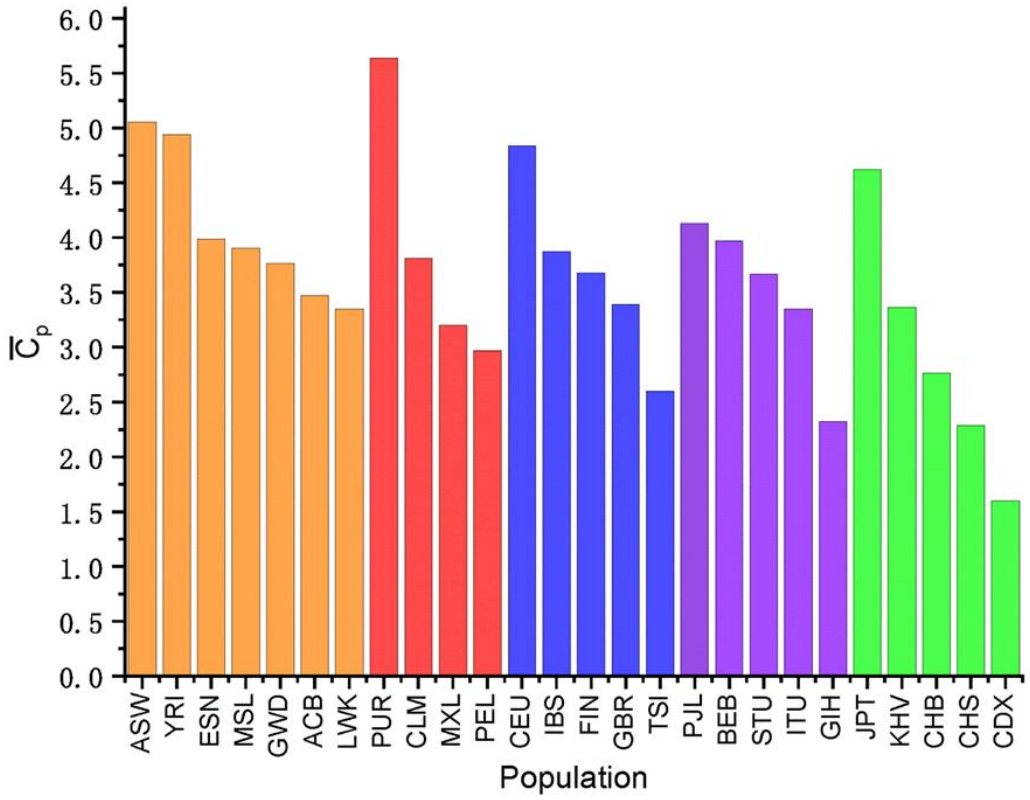

Figure 2.13 - Average count of MIs per individual among 26 populations. The average ranged from 1.60 to 5.64.

The MI counts in each individual among the five super-populations, which were used for calculating the means and standard errors of MI numbers in five super-populations are the number of MIs in each individual. Notably, Africa had the highest count of MIs per individual, 3.96, and East Asia had the lowest, 3.03. This result is consistent with the recent studies of the human genetic SNP analysis, in which Africa has the highest SNP number and East Asia has

the lowest SNP number. The average counts of MIs per individual for the 26 populations classified by five super-populations, are also shown in Figure 2.13.

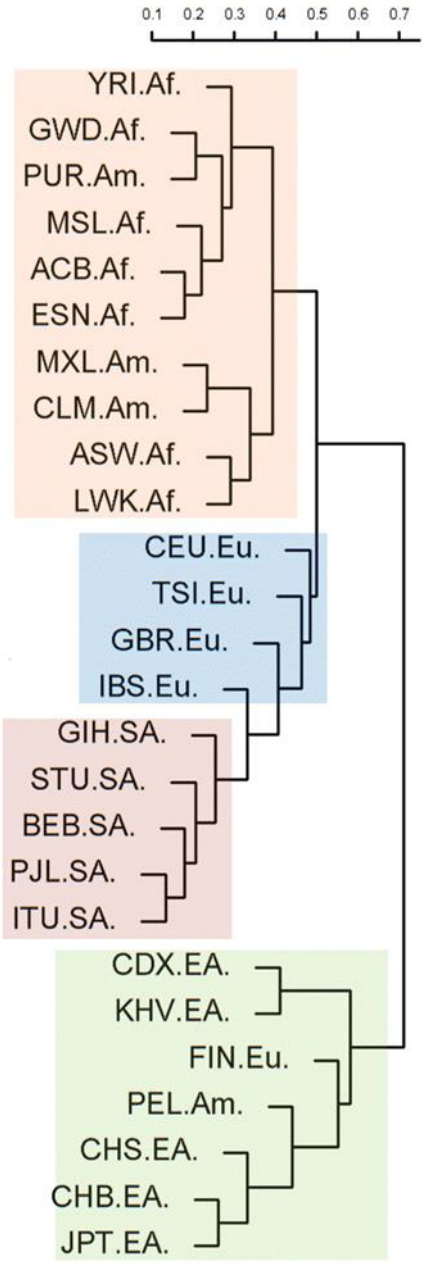

Figure 2.14 - Phylogenetic trees for the 26 populations based on the MIs.

Since the average MI count is 3.6 (varied from 1.60 to 5.64, see Figure 2.13) for all 26 populations, we found that the group with the average MI count over 4.5 consisted of five populations: PUR (Puerto Ricans from Puerto Rico, Cp = 5.64), ASW (Americans of African Ancestry in SW USA, Cp = 5.05), YRI (Yoruba in Ibadan, Nigeria, Cp = 4.94), CEU (Utah Residents (CEPH) with Northern and Western Ancestry, Cp = 4.84), and JPT (Japanese in Tokyo, Japan, Cp = 4.62). We noted that the top two populations with high-frequency MIs, PUR (Cp = 5.64) and ASW (Cp = 5.05), also have a high degree of admixture in the analysis of SNP analysis in the 1KGP study. This suggests a potential influence of ancestral lineage mixture on MI counts in populations. We will discuss the five populations in details in the section of Discussion and Conclusions.

In order to examine the genetic structure of MIs hierarchically among populations, we constructed a genetic distance-based phylogenetic tree using all of the MIs detected from the 26 populations (Figure 2.14), which yielded results that generally provide evidence of genetic clusters: African and American populations were clustered into one branch, European and South Asian populations into a second, and East Asian populations into a third.

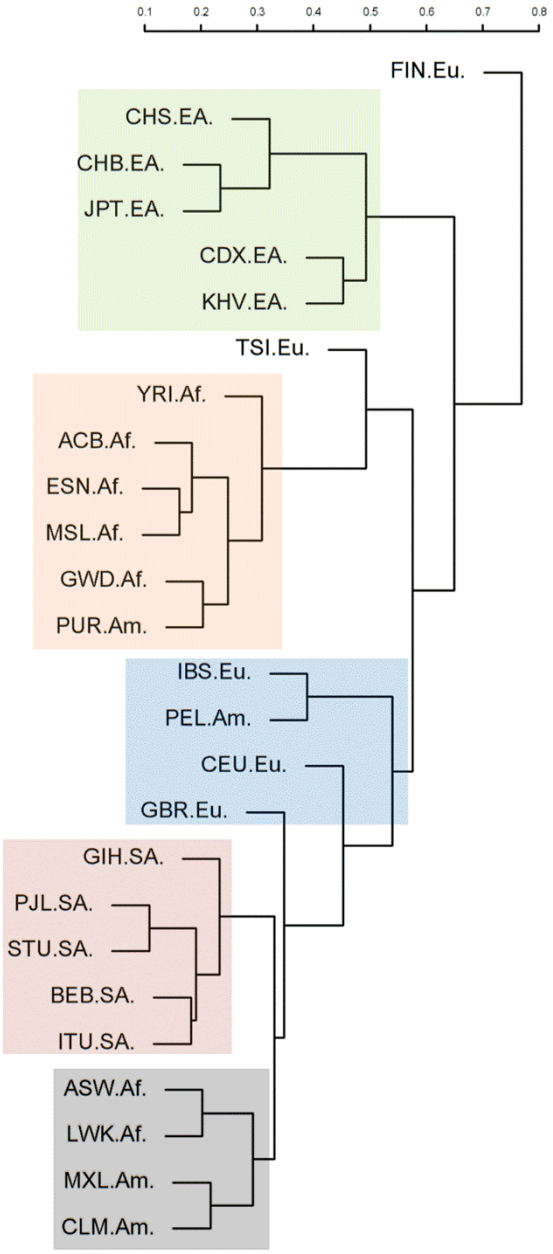

Figure 2.15 - Phylogenetic trees for the 26 populations based on the MIs in gene regions. The average pairwise population distance was computed with the 2,135 MIs in genetic regions, constructed by neighbor-joining.

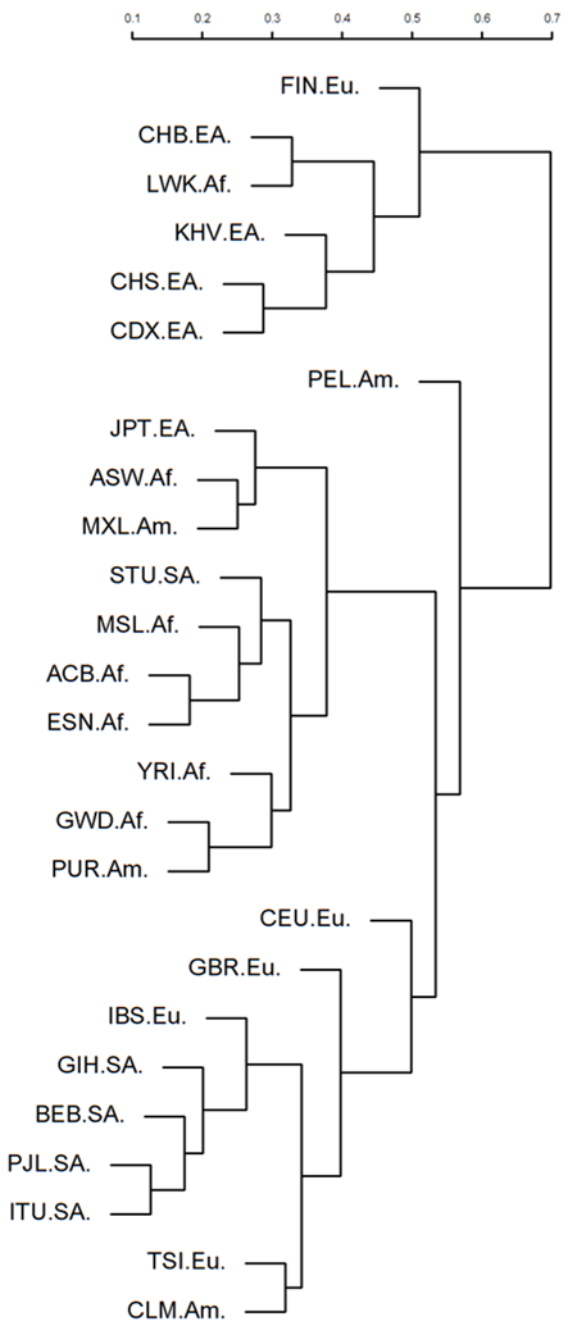

Figure 2.16 - Phylogenetic trees for the 26 populations based on the MIs in intergenic regions.

2.3.3 Population structure based on MI statistics

We also performed phylogenetic analysis with MIs in gene regions only to construct a phylogenetic tree (Figure 2.15). Similar to the full MI analysis in Figure 2.15, the phylogenetic analysis with MIs in gene regions in Figure 2.15 displayed similar cluster into four branches. Besides, we used MIs in intergenic regions only to construct a phylogenetic tree (Figure 2.16). As shown in Figure 2.15 and 2.16, the populations were more dispersed in phylogenetic tree with MIs in intergenic regions compared with that in gene regions.

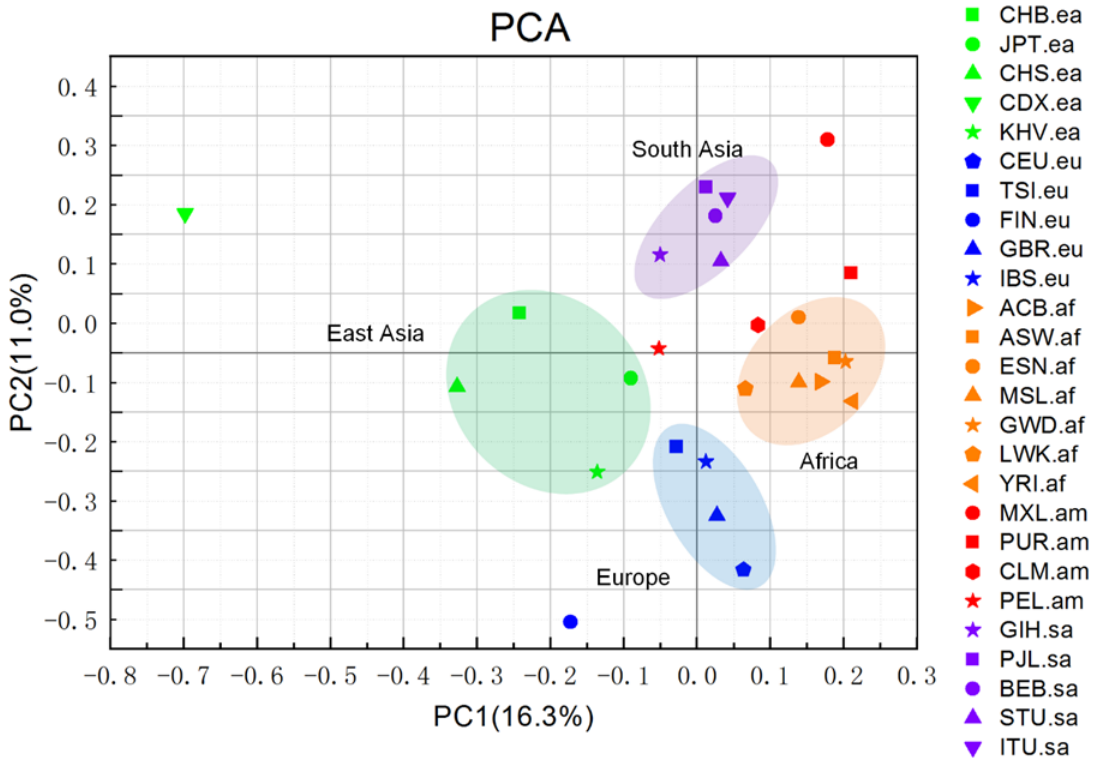

Figure 2.17 - PCA of 26 populations. The results are plotted as the first and second principal components. Different color fills represent the five super-populations: Orange, Africa; red America; blue Europe; purple, South Asia; green, East Asia..

The topology of the phylogenetic tree not only met our expectation but also reflected a pattern among the 26 populations suggesting that ethnic groups that live geographically closest to one another have a relatively small MI genetic distance. An exception was the FIN (Finnish

in Finland) population (Clustered with the East Asian populations in Figure 2.14 and formed a single cluster in Figure 2.15), which deviated from the Europe branch. This may possibly result from Finland's unique language linkage. Not like other Europeans, Finns speak kind of Uralic language instead of an Indo-European language. The phylogenetic tree revealed that MIs with functions similar to those of SNPs are key to tracing the evolution of the genetic structure of human populations.

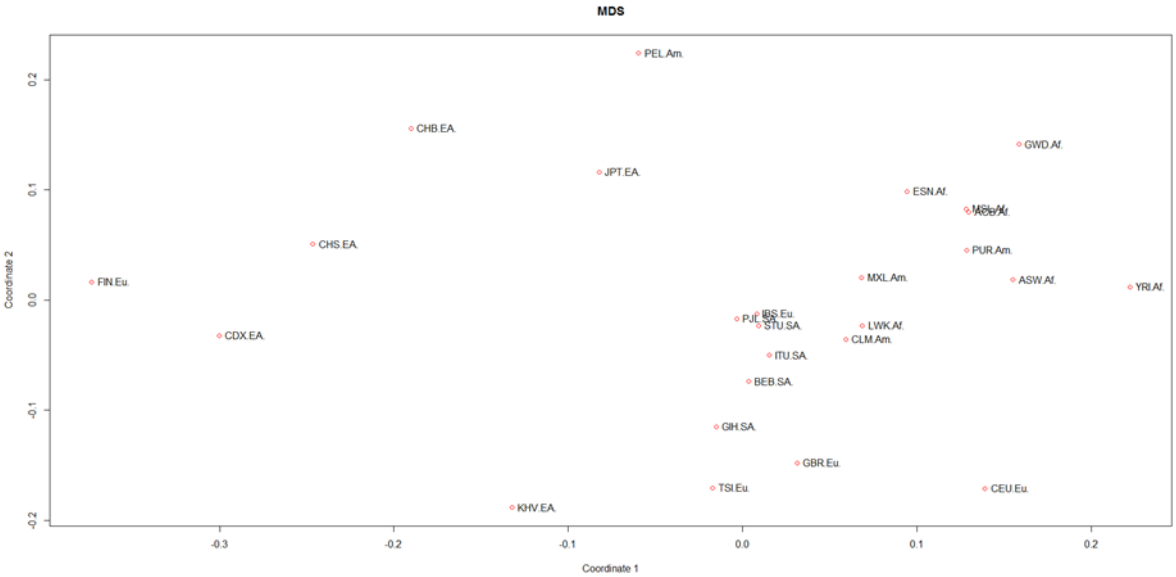

Figure 2.18 - MDS of 26 populations based on all the 6,968 MIs.

To further expose the undiscovered relationships of the MIs among the 26 populations, we performed principal component analysis (PCA) of the 26 populations with all the 6,968 MIs. As shown in Figure 2.17, the PCA of MI patterns revealed that the 26 populations were divided into four groups according to the top two main components. This result indicated that the populations in same super-population were closer to each other. We found that African, European, South Asian and East Asian populations are clearly recognizable, which

represented genetic drift through human evolution or other factors. Unlike the other four super-populations, the American populations are dispersed and do not form a distinct cluster. Our results are well consistent with the widespread pattern from SV PCA analysis. We also performed multidimensional scaling (MDS) analysis with all the 6,968 MIs displayed in Figure 2.18. Similar to the results in the PCA, the MDS result also shows that the populations in the same super-populations are closer compared with those in different super-populations.

2.3.4 MIR sharing analysis among five super-populations

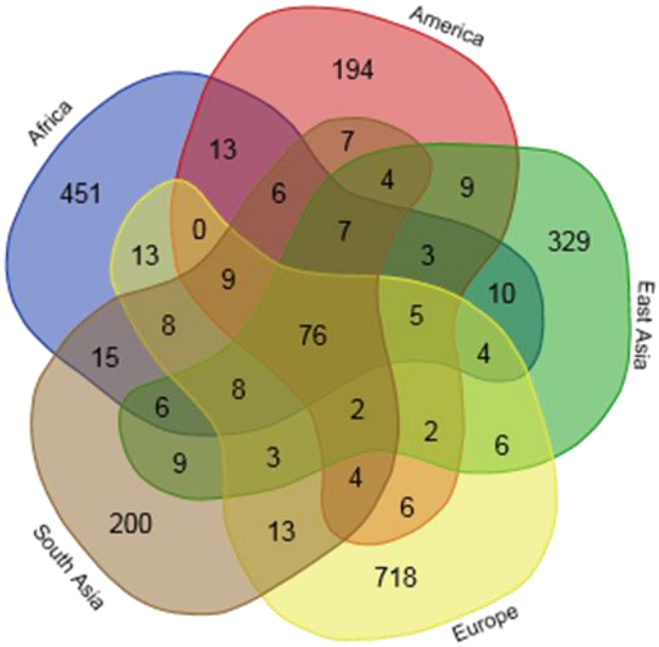

Figure 2.19 - Venn diagram of all MIRs sharing results among the five super-populations.

Since the distinct variation pattern of each population may implicitly affect phenotype divergence among the population, it is significant to explore the common and distinct MIRs among the five super-populations to see if there are any large difference of MIs among the

five super-populations. To further investigate the diversity and relationship of MI among the five super populations, we conducted an MIR sharing analysis with a Venn diagram by counting all the MIRs (Figure 2.19). We found that 76 MIRs were shared by all five super-populations.

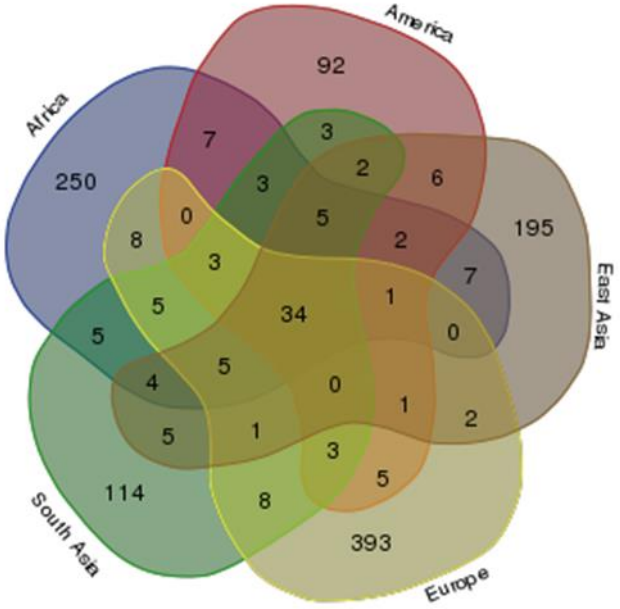

Figure 2.20 - Venn diagram of MIR sharing results in only the gene regions according to the GENCODE database among the five super-populations.

Furthermore, we used the MIRs located only in the genetic region in the Venn diagram (Figure 2.20), 34 MIRs within gene regions according to the GENCODE database were shared by all five super-populations (Figure 2.20). In addition, no MIR in the gene region was shared by the four super-populations, America, Europe, South Asia, and East Asia except the 34 MIRs that are shared by all five super-populations. However, any four of the super-populations, as long as Africa was included, shared at least one MIR in addition to the 34. This result suggests that any one of the four super-populations America, Europe, South Asia, and East Asia is

closer to African than to each other. Venn diagram also indicates that some MIs are shared among the five super-populations and the other MIs are private to one specific super-population or private to a population. In general, each population has private MIs and common MIs shared with other populations in the same super-population. This result is also consistent with the SV polymorphism in the previous study.

In the current study, we define "MIR hit" for an MIR as the number of MIs included within the MIR. Consequently most MIRs in gene regions were unique in every super-population (i.e., singleton MIRs), especially for the Europe population. There was a total of 1,008 singleton MIRs and 36 non-singleton MIRs. The full list of MIR hit counts in gene region among super-populations appears in Table 2.3. According to the results listed in Table 2.3, the most recurrent non-singleton MIR in gene region had a hit of 8, and was in the super-population of Africa. The next most recurrent MIR (hit = 7) was also in Africa. This may be due to the fact that Africa consisted of multi-ethnic groups.

Table 2.3 - MIR hit counts in gene region among super-populations.

Super-populations	1 hit count	2 hit counts	3 hit counts	4 hit counts	5 hit counts	6 hit counts	7 hit counts	8 hit counts
East Asia	192	1	1	1	0	0	0	0
Europe	388	4	0	1	0	0	0	0
Africa	225	12	6	2	2	0	2	1
America	91	1	0	0	0	0	0	0
South Asia	112	2	0	0	0	0	0	0

2.3.5 Effects of MIs on human health

Table 2.4 - Population specific Genes overlapped with MIRs that hit is over three.

Gene	Super-populations	Hit	Influence on Health	MIR Annotation	Chromosome
RP11-325I22.2	Africa	8	Lung adenocarcinoma	Intron	Chr5
ANKRD36	Africa	7	Maximum Number of Alcoholic Drinks Americans.	Intron	Chr2
				Exon (3' UTR)	Chr2
SVEP1	Africa	7	Hirschsprung's disease; Coronary Disease	Intron	Chr9
HBG2, HBE1, AC104389.28	Africa	5	Hemoglobinopathy toms river and cyanosis, transient neonatal.	Intron	Chr11
PRPSAP2	Africa	4	Osteosarcoma	Intron	Chr17
FBL	Africa	4	Scleroderma and myositis.	Intron	Chr19
ANK3	East Asia	4	Schizophrenia; bipolar disorder;	Intron	Chr10
PYY	Europe	4	Obesity	Intron	Chr17

Although the MIs in healthy individuals have not made an immediate impact or caused diseases directly at present, we infer that MI occurrences in gene regions rather than the intergenic regions are more likely to affect gene function, and thus potentially result in the phenotypic diversity of disease susceptibility among various populations in future. To investigate the potential association of MIs with gene function and human diseases among various populations, we conducted the analysis of the genes that are frequently affected by MIs. We focus on eight genes that overlapped with MIRs (Table 2.4), with hit >3 listed in

Table 2.3. We obtained the gene function through GeneCard (http://www.genecards.org/) as well as related reference studies and listed them in Table 2.4. As shown in Table 2.4, MIs do cause substitutions against the same position on the human reference genome hg19 [116] in the gene regions. Although it is not clear how these MIs in intron or exon (UTR) regions affect the functions of genes, it is reported that both UTRs and exon regions play an important role in translation and transcription [125,126].

Table 2.5 Count of common MIs among seven non-human primates and five human super-populations.

Name	All MIs	East Asia	South Asia	Europe	Africa	America	1KGP
Chimpanzee (panTro4)	2,284	12	11	7	14	11	20
Gorilla (gorGor3)	2,010	10	11	9	14	10	18
Orangutan (ponAbe2)	3,473	9	9	7	10	9	14
Gibbon (nomLeu1)	3,713	6	7	5	7	3	13
Baboon (PapHam1)	4,547	5	5	3	6	5	8
Rhesus (rheMac3)	4,271	5	4	2	5	4	7
Squirrel monkey (SaiBol1)	4,214	2	2	2	2	2	2
Total	24,476	49	49	35	58	44	82

In addition, most of these genes listed in Table 2.3, are reported to be related with health. Among these genes, one named ANKRD, overlapped with the MIs in only African populations. This gene is associated with genes that control the capacity to tolerate alcohol [127]. In addition,

we found gene ANK3, which overlapped with MIs in only East Asians, have been reported previously to be a risk factor for schizophrenia, a chronic and severe mental disorder, in Han Chinese[128], which may explain the only existence of MIs in ANK3 in East Asians.

Furthermore, we compared the MIs in human with the seven closely related non-human primates (chimpanzee, gorilla, orangutan, gibbon, baboon, rhesus monkey, and squirrel monkey) to investigate the inheritance and evolution of MIs. By comparing MIs in the genomes of non-human primates against the hg19 [116] human reference genome and MIs detected in the genomes of the 1KGP super-populations, we discovered a larger number of MIs in Africa overlapping with the seven non-human primates. As shown in Table 2.5, the number of MIs shared by non-human primates and five super-populations are in the descending order: 58 shared with African, 49 with East Asian, 49 with South Asian, 35 with European. The higher overlapping ratio of MIs between the non-human primates and Africans might be due to that African inherited these shared MIs from the common ancestor of non-human primates and human. Indeed, we inferred that after the Out-of-Africa geographic movement of modern human, Africa ancestral population would have spread the distribution of the most MIs, and partial of these MIs would survive in contemporary populations, resulting in the smaller MI number among the other four super-populations. However, further sophisticated fossil and genetic data are necessary and expected to confirm this inference.

2.4 Discussion

As we analyzed the MI count per individual among 26 populations, we found that PUR (Puerto Ricans from Puerto Rico) and ASW (Americans of African Ancestry in SW USA) have the most MIs per individual. These two populations with high-frequency MIs also have a high degree of admixture in the previous SNP study [115]. This suggests that admixture degree may potentially affect the MI count by multiple ancestral lineages among populations, which were also reported to affect genetic diversity by previous studies [129]. Specifically, ASW genome has been reported to come from three ancestral lineages with 75.9% of African, 21.3% of European, and 2.8% of Native American [130]. PUR, with a particular geographical stratification along the island and the ancestry of Amerinds, Spanish, and Africans, has the potential to enlarge the expected MI diversity [131,132]. As for the rest three populations with high-level MIs including YRI, CEU and JPT, the reasons for high-level of MIs may vary with more complex implications. We noticed that YRI was distinguished from other Africans with significant heterogeneity when the population structure was compared with other African populations in another study [133]. Although the admixture degree result showed that the admixture degree of CEU was not as high as PUR and ASW, it was reported that CEU descended from migrants originating from northern and western parts of Europe [134]. With the lowest MI count among the five populations, JPT was reported a noteworthy proportion of the special Jomon ancestry in the modern Japanese, which leads to JPT's genetic features quite distinct from other East Asian populations although the admixture degree is not as high as other four populations here[135,136]. However, we believe that the further sophisticated analyses

of population structure and migration history are necessary to confirm the reasons for the high-level of MIs in these five populations.

In the phylogenetic analysis of the 26 populations, geographical distributions and migrations affected the MIs. African and American populations were clustered into a branch on the phylogenetic tree perhaps because African Americans are the largest racial minority [137] and a mixed heritage is common in America. Specifically, the populations on the East Coast of America and in Western Africa were clustered on another branch. This cluster may be due to the black slave trade from Western Africa to the East Coast of America [138]. The populations on the West Coast of America and in Eastern Africa were closer than other populations. This closeness is consistent with the migration of people from Eastern Africa to the West Coast of America [139]. We found that European and South Asian populations were clustered into one branch in Figure 2.16, which has also been reported in other studies focusing on copy number variations and retro-duplications [140]. The specificity of the Finnish population (FIN) can be interpreted in terms of multiple genetic components and demographic factors such as isolation, migration, and admixture, which are reflected in their distinctive distribution among the European populations in another study [141]. Moreover, our finding that PUR (Puerto Ricans from Puerto Rico) are closer to populations in East Asia is reasonable because some Peruvians are Asian immigrants [142].These MIs used in phylogenetic tree and PCA analysis could also be used to cluster different super-populations, showing that these MIs represented different genomic backgrounds of the studied populations. The phylogenetic analysis of the 26

populations are consistent with the corresponding migration history and ethnicity composition of these populations.

We arranged the five super-populations in descending order of average number of MIs per individual as follows: Africa > America > Europe > South Asia > East Asia. Among these five super-populations, the average numbers of MIs per individual supported the "Out of Africa" hypothesis, which believed that humans originated in Africa, then recent Africans migrated to other continents. Scientists speculated that modern Africans first migrated to Eurasia, which is the continents of Europe and Asia considered together. In addition, the blood genetic distance of Africans and Europeans is smaller than that of Africans and Asians, which implies a closer relationship of Africa and Europe [143]. Specifically, as reported by Macaulay *et al* [144], the initial branch of the Asians from the southern dissemination pursued the Nile from the east of Africa, went towards north and aimed to get into Asia. When through the Sinai, the crowd branched, some shifted into Europe and the others went on heading into Asia. This assumption is based on the comparatively late date of the landing of modern humans in Europe. Through the process to East Asia, a small part of the African migrated along the Arabian Peninsula and India Cost in South Asia, and arrived at Australia finally [143, 144]. The high MI count in America is possibly be due to the African ancestry of these American populations and its complex population structure.

It is known that long-term migration can cause genomic variations [145]. We inferred that after the Out-of-Africa migration of modern human ancestors and hybridization among the

four non-African super-populations, the Africans introduced some MIs into the ancestors of the other four super-populations. It is speculated that the differential coevolution of MI lineages with different but closely-related ancestral populations and subsequent MIs in parallel with the introgression of archaic alleles into the genomes of modern human ancestors may be largely responsible for the present-day variant counts of MIs in multiple populations. Therefore, our results may be evidence of the "Out of Africa" hypothesis.

Evidence of the "Out of Africa" hypothesis also stems from a comparison between the MIs of non-human primates and those of humans. Africans shared the most MIs with the seven non-human primates, but the other four super-populations only shared a few. The higher overlapping ratio of MIs between the non-human primates and Africans might be due to that African inherited these shared MIs from the common ancestor of non-human primates and human in the ancient time. It may be assumed that, after the Out-of-Africa geographic movement of modern human, Africa ancestral population would have spread the distribution of the MIs which are shared with the seven non-human primates, and partial of these MIs would survive in contemporary populations. Indeed, the periodic genetic interchange among human super-populations, both in ways of frequent gene flow constrained by geographical distance and of population extension events causing interbreeding[On account of that not all of the MIs in African will be exchanged into the other four super-populations, smaller number of MIs survived among the other four super-populations. However, further sophisticated fossil and genetic data are expected to confirm this inference. We found MIs in three genes, *MYH14*,

GRM7, and *DFNA5,* all related to hearing[146-148] in both humans and non-human primates. This suggests that the MIs in these three genes may have existed in the common ancestors of humans and non-human primates and were conserved in some non-human primates and some human super-populations. Therefore, it may be that the super-populations that inherited the MIs may differ from others in their sense of hearing.

The analysis also showed that some of the regions in the human reference genome hg19 do not include complete information. Of the 76 MIs in the five super-populations, 34 overlapped with genes, and 35 were also found in the seven non-human primates. Compared with the most recent common ancestor of the human population, hg19 may be inverted in the regions that overlap with the 34 MIs because such regions are conserved among close species. However, some of these regions have been reported as single-nucleotide variations because of the limited understanding of MIs. In addition, hg19 has been reported to contain rare alleles, according to the Genome Reference Consortium [149]. Accordingly, the reference of human MIs should be taken into account in updates of the reference of human assembly in the future studies.

Here, we focused more on the effects of MIs on human evolution. However, as we analyzed the genes that overlapped with MIRs, we found that MIs may be associated with human health. Though the functions of the intron and UTR regions affected by MIs with limited samples in this dissertation are not clear yet by now, these MIs may potentially benefit studies of relation of variation and health in the future. Besides, the MIs shared only by one

super-population, especially those overlapped with gene regions, may increase the susceptibility of some diseases. With limited samples from healthy individuals in the 1KGP, MIs in this dissertation are possible candidates for future studies. Accordingly, more MIs from samples of diseases are needed to explore the correlation of MIs and human health. Nevertheless, we expect personalized medical information in the future can help us better understand the relationship of MIs and health, as well as unveil of disease mechanisms.

2.5 Conclusions

In this study, we made a comprehensive analysis of MIs on non-disease individuals from the 1KGP and seven non-human primate genomes, then built a landscape of human populations and explored the effect of MIs on human diversity, evolution and health. Applying the software MID[43]. for 26 human populations and searchUMI for alignments between the human and seven non-human primate genomes, respectively, we analyzed the 6,968 MIs detected in 1,937 individuals of the 26 populations from the 1KGP and the common 82 MIs among the seven non-human primate genomes and the 1,937 individuals. It is shown that the MIs from five super-populations (Africa, America, Europe, South Asia, and East Asia) reveal population structures at multiple levels and may affect individuals in several aspects. Firstly, the widespread MIs in human genomes greatly contribute to human diversity, which could be a result of geographical distribution and human migration. Secondly, the large-scale MIs are good supplements to existing markers in reconstructing human evolutionary history. Thirdly,

the MIs in particular super-populations, especially those overlapped with gene regions, are informative for understanding the association of health and genetic variations.

To our knowledge, our study is the first to analyze MIs in humans on the population scale. In this dissertation, MIs are displayed in many levels including short reads, individual genomes, population-scale, and ancestry-scale. Furthermore, the MIRs supported by multiple short reads, individuals genomes, various populations will contribute to the concrete understanding of MIs. Our analysis of MIs with the 1KGP data improves our understanding of human genetic diversity and evolution. The comparative analysis of MIs at the population, super-population, and species scales are thus expected to contribute to further implementation of human evolutionary theory. Future large-scale sophisticated fossil data and archaeological materials for analyzing the age of MI polymorphism will be informative for understanding the behavior as genetic markers and better reconstructing the human evolution history.

CHAPTER 3 ANALYSIS OF MICRO-INVERSIONS IN CANCER GENOMES

3.1 Introduction

Many SV patterns have been observed by studies on cancer genomes. Besides, SV detection tools based on the high-throughput sequencing data permit researchers to single out the patterns of structural variations that characterize special type of cancers in recent years. These studies reflect that SVs are not restricted to one type cancer. Besides, studies show that the genomes of different cancers display various prevalence of variations on chromosomes, and regions of the genomes. Though many SVs don't directly cause cancers, some could contribute to serious illness as a result of changing the function of regulatory regions such as enhancer elements and promoters [cl]. Other SVs may affect cancers by influencing the protein signal pathways.

A few studies show that the cancer individuals of one type of cancer all have some SVs in unique genes. For example, two genes DICER1 and DROSHA are frequently found with copy number variations in non-small-cell lung cancer patients [cli]; the stem cell marker, SALL4 were often discovered in Asians with hepatocellular carcinoma [clii]; Research shows that most pancreatic cancer patients have the somatic mutation on the gene KRAS [cliii]; deletion of ETV6/RUNX1 gene was found in patients with acute lymphoblastic leukemia [cliv]; Reproducible copy number variations in MAPKAPK2 promoter were proved to have impact on lung cancer patients [clv]; and the gene DEFB4 in each sample with cervical cancer in the study of Abe et al. was identified with copy number variation[clvi].

In consideration of that SVs in various types of cancers are likely to have distinct preferences of discrete gene regions, analyzing the genes with some frequent variations among numerous cancers is exclusively significant. In addition to the different SVs among various types of cancers, some common regions on the genome were found with highly-unified SVs in many cancers. For example, it is reported that tumors were found with identical replicating sequences on the gnome [clvii,clviii]. In general, some variations occur in only one special type of cancer, but the other SVs are in common by multiple cancers.

Although many studies have focused on the association of SVs and cancers previously, MIs have never been included in these analysis. Therefore, the research field of MIs in cancer genomes are still blank. In this part, we have conducted a thorough analysis of MIs in six cancers including 145 samples of esophageal cancer, 131 samples of bladder cancer, 67 samples of hepatocellular carcinoma, 62 samples of lung cancer, 32 samples of prostate cancer, and 14 samples of pancreatic cancer from SRA database.

3.2 Materials and methods

The MI analysis in this part mainly includes collecting samples, detecting and annotating MIs, analyzing the chromosomes that MIs prefer to be located in, analyzing the top genes MIs frequently happen in each cancer and comparing MIs with SNPs which overlapped with MIs in the same genome regions. The pipeline for the whole analysis is displayed in Figure 3.1. The details of the materials and methods are described as follows:

3.2.1 Sample collection

Besides the databases storing healthy individuals such as 1KGP, cancer genomic database also exist extensively such as Sequence Read Archive (SRA) database [19]. A comprehensive understanding of genomic variation and regulation of cancer could lead to a better understanding of cancer mechanisms and genomic manifestations [clix]. The analysis of variations in cancer genomes are helpful to drug sensitivity predictions based on heredity, lineage, and gene expression. Researches have shown that large-scale annotated aggregation of cancer sample information may contribute to clinical validation of anticancer agents, clinical prediction of drug response and the design of related cancer treatment plans, which may accelerate the pace of personalized cancer medicine [clx].

SRA database stores raw sequencing data and alignment information based on high-throughput sequencing platforms, including Roche 454 GS System®, Illumina Genome Analyzer, Applied Biosystems SOLiD System, Helicos Heliscope, Complete Genomics, and Pacific Biosciences SMRT. One of the advantages of SRA is that it stores sequencing data not only healthy individuals but also cancer samples.

The high-throughput sequencing data in SRA mainly includes two categories: whole genome sequencing (WGS) and whole exome sequencing (WES). Generally, WGS sequence the complete DNA of the whole genome at a single time. This process brings about sequencing all of the genome chromosomal DNA as well as DNA contained in the mitochondria and, for plants, in the chloroplast. Thus, whole genome sequencing is the most comprehensive approach to genome research. Genomic information could be used to identify genetic diseases, find mutations that drive cancer development and track disease outbreaks.

The rapidly falling cost of sequencing and the increased ability to process large samples of data have made whole-genome sequencing the most powerful tool available to today's sequencers. However, WGS has its own limitation. For example, WGS is time consuming and the cost of WGS is really high compared with whole exome sequencing (WXS). Besides, since the WGS needs to cover all the sites on the genome, the sequencing depth is usually not high. WGS is often understood to be used to determine the human genome, but the scale and flexibility of high-throughput sequencing is such that it can be used efficiently in any species, such as animal husbandry, plants, disease-related microorganisms and non-human primates. In this part, the data used for analyzing cancer genomes are all WGS.

It's important to depend how cancer genome data be used for cancer biology and cancer therapy. With the advent of high-throughput sequencing technology, it's possible to make comparative genomics analysis between clinical cancer patient and personalized individuals without diseases [clxi]. Indeed, people can better understand cancer genome characteristics, cancer causes and possible mechanisms.

In this dissertation, we collected BAM files of the unmapped reads of 451 samples from SRA database. These samples are patients of 145 from esophageal cancer, 131 from bladder cancer, 67 from hepatocellular carcinoma, 62 from lung cancer, 32 from prostate cancer, 14 from pancreatic cancer. All the 145 Illumina samples are pair-end sequencing reads based on WGS. Each of these short reads is 151 bp. In addition, the samples for this part include both males and females with age of 51 to 70 years old. Besides, we adopted the 1,937 normal individuals from the third phase of 1KGP, which were described in the

part of 2.1, as the control samples. These data allow for the comprehensive comparison of health cancer samples and healthy samples as well as samples of the six different cancers.

3.2.2 MI detection and annotation

Identical to detecting MIs from the individual genomes from 1KGP, we applied the tool MID [43] to identify MIs in the individuals of the six cancers. MID was set with default parameters: the anchor length was set at 18 bp, and the mismatches length was set at 2 bp. Then, MID remapped the total unmapped short reads to the human reference genome hg19. MID is a reliable tool proposed by our previous work, which is able to intensively detect MIs from bam files containing unmapped short reads with dynamic programming algorithm. Furthermore, MID is extremely perceptive to fine-scale MIs of which length are 10 to 100 bp. Although some of the SV detection tools proposed in previous studies can also detect inversions, the length of the detected inversions are too large usually larger than one thousand base pairs, which are not comparable to the length of MIs (10 to 100 bp).

After detecting MIs with MID, we annotated MIs with annotation files from GENCODE database based on hg19 coordinates. GENCODE is a genetic annotation database that identifies genetic characteristics through a series of computational analyses, manual annotations, and experimental results. GENCODE integrates a lot of gene information such as protein coding, long noncoding RNA (lncRNA), coding sequence (CDS) and is widely used in genome annotation. During this process of gene annotation in cancers, MIs were annotated as intergenic and genetic. In case that an MI was annotated as intergenic, it means this MI was in intergenic region and didn’t overlap with any gene

regions; if an MI was annotated as gene, it means this MI was overlapped with at least one gene of the human assemble hg19.

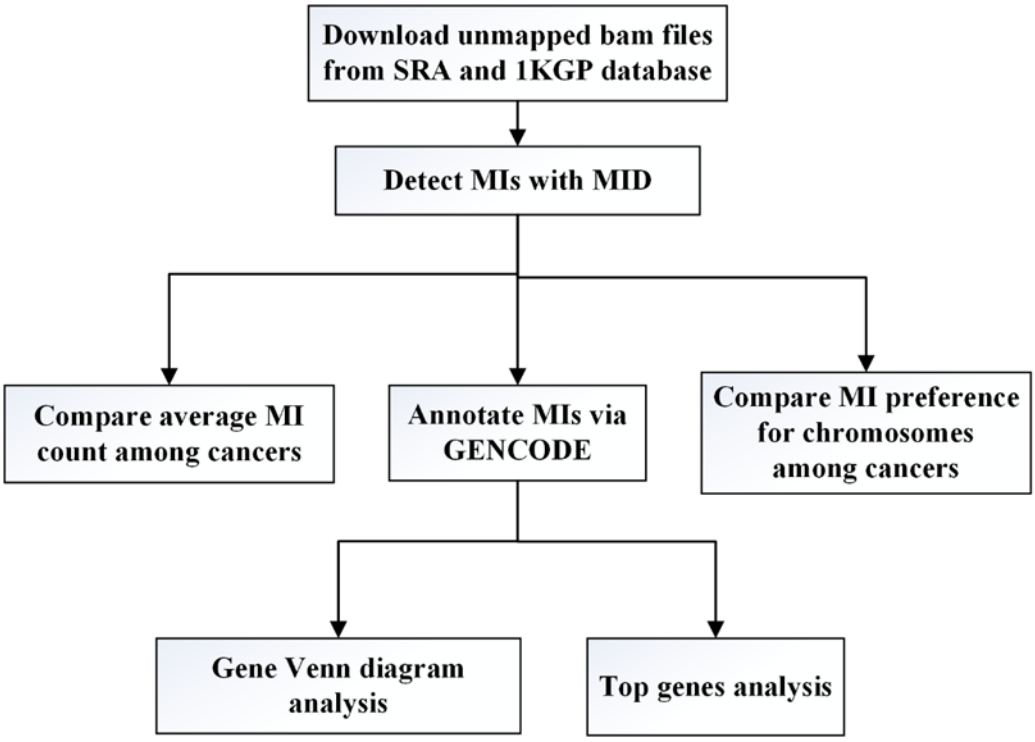

Figure 3.1 - The pipeline of the analysis of MIs in cancers.

3.2.3 Average MI count per Individual

Considering that the sample number is different in the six cancers, the MI characteristics of the cancer genomes and the normal samples will be better displayed through the average count of MIs per individual in one cancer than the total MI number of each cancer or healthy group. To address this point, we define parameter $\overline{C_p}$ to represent the average count of MIs per individual among cancer samples or healthy samples as following:

$$\overline{C_p} = \frac{\sum_{i=1}^{N_p} C_i}{N_p}, \quad (1)$$

In this evaluation, C_i means the count of MIs per individual, and N_p means the total number of individuals in a cancer or the healthy group from 1KGP.

3.2.4 Venn diagram analysis

To further interrogate the diverse patterns of special genes overlapped with MIs among the six cancers, we performed a gene sharing analysis by counting all the genes overlapped with MIs in each cancer. Since the Venn diagram is only available for at most five categories, we removed pancreatic cancer with the least sample size of only 14 and included the rest five cancers including esophageal cancer, bladder cancer, prostate cancer, lung cancer, and hepatocellular carcinoma in this analysis. The process of conducting the Venn diagram is in a few steps. Firstly, we selected the MIs that were annotated as gene region as reported by the annotation files from GENCODE database. Then, we extracted the genes overlapped with these MIs in each of the five cancers. Next, we filtered these genes, removed the repeated genes, and obtained the unique genes. At last, we applied these unique genes to perform a Venn diagram to discover the shared common genes of the five cancers.

3.2.5 Comparison with SNPs

In consideration of the short length of MIs, we speculated that MI sites on the human genomes have been improperly reported by the previous studies as SNPs. Thus, we have downloaded the list of the exact sites of mutations in genes annotated in the Catalogue Of

Somatic Mutations In Cancer (COSMIC) database [clxii], the largest somatic mutation depository to date, to see whether there is an overlap between MIs analyzed in this dissertation and the SNPs reported before. We got all the COSMIC coding point mutations from both targeted and genome wide screens from the recent release (v90). On account of that the SNPs from COSMIC are coordinated with the human reference genome hg38, we converted the genome coordinates from hg38 to hg 19 with the tool CrossMap [clxiii] with file option of vcf.

3.3 Results

In this part, we analyzed 451 samples from SRA database comprising six types of cancer including 145 samples from esophageal cancer, 131 samples from bladder cancer, 67 samples from hepatocellular carcinoma, 62 samples from lung cancer, 32 samples from prostate cancer, and 14 samples from pancreatic cancer. With MID tool, we detected a total of 12,532 MIs, of which 5,995 (47.84%) overlapped with gene regions. Of the total 12,532 MIs, 3,917 MIs were unique. In other words, the rest 8,615 MIs were repetitive to at least another MI in the 12,532 MIs. Besides, we included the 1,937 normal individual genomes from 1KGP as the control samples. As reported in the section of 1KGP analysis, which included healthy people genomes, we identified 6,968 MIs, of which 3,549 (50.93%) MIs were in gene regions. Of the 3,549 MIs in gene regions, only 370 (10.42%) were in exon regions. The comparison of MI distributions of cancer samples and healthy individuals implies that the possibility of MI occurrence in cancer genomes (12,532 MIs/451 samples) is much higher than that in healthy samples (6,968 MIs/1937 samples). However, the frequency of MIs in gene regions of both the cancer samples (47.84%) and healthy samples

(50.93%) is high. The details of MIs detected in the cancer genomes are shown in Table 3.1. With all the samples detected in this part, we do the following MI analysis.

Table 3.1 - Summary of MI results.

Cancer Type	Sample number	MI indicator			
		MI-num[a]	MI-unique[b]	MI-gen[c]	MI-ave[d]
Esophageal cancer	145	4,873	1,042	2,224	33.61
Bladder cancer	131	4,103	1,528	1,977	31.32
Hepatocellular carcinoma	67	359	304	208	5.36
Lung cancer	62	977	976	519	15.76
Prostate cancer	32	2,082	791	985	65.06
Pancreatic cancer	14	138	118	82	9.86
Total	451	12,532	3,917	5,995	27.79
Healthy Samples	1,937	6,968	2,140	3,549	3.60

a. The "MI-num" column illustrates the number of MIs for each cancer.
b. The "MI-unique" column illustrates the number of unique MIs that are not overlapped with each other.
c. The "MI-gen" column illustrates the number of MIs that overlap with genes for each cancer.
d. The "MI-Ave" column illustrates the average number of MIs per individual in each cancer.

3.3.1 MI distribution on chromosomes

The MI distribution on human chromosomes in all the cancer genomes and the healthy genomes is shown in Figure 3.2. It is well known that human chromosomes are numbered in the decreasing order of chromosome length. For example, Chromosome 1 has the most base pairs and are the longest of all the 24 chromosomes. As shown in Figure 3.2, the MI count generally increased with the chromosome length increasing. This result is logical due to that longer chromosomes have more base pairs and more chance to make a mistake in the DNA replication. Thus, the MI occurrence frequency in longer chromosomes may increase. However, the MI number is not strictly linearly dependent to the chromosome

length. For example, the result in the Figure 3.2 shows that Chromosome 3, 6, 13, 19 in cancers performs differently with other chromosomes.

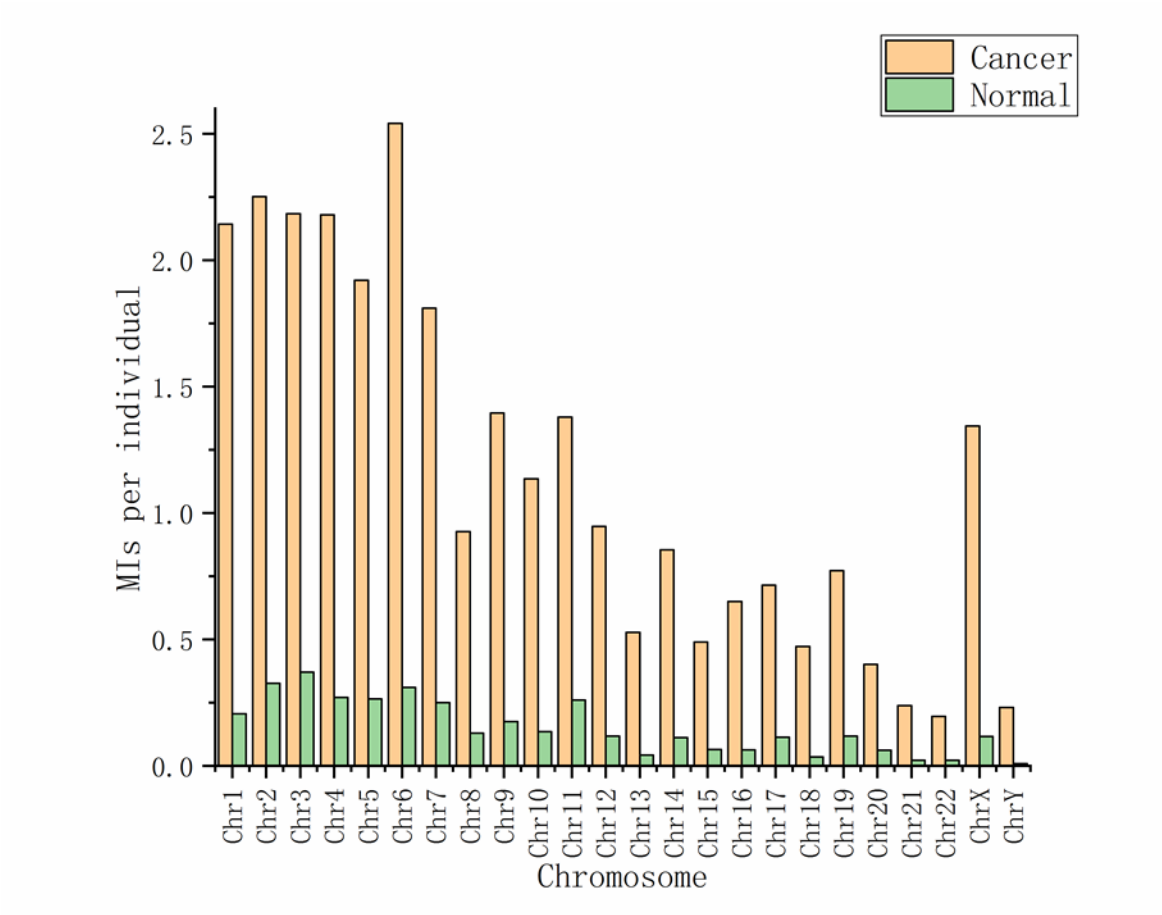

Figure 3.2 - MI distribution among 24 chromosomes in cancer and healthy samples. The yellow bars represent MIs in cancer samples and the green bars represent MIs in normal individuals.

Specifically, the MI number of Chromosome 3 is higher than that in Chromosome 2, while the Chromosome 2 have more base pairs than Chromosome 3. Chromosome 6, 13, and 19 have similar situations. Different with cancer genomes, the special chromosomes are 19, 17, 16, and 11 for the healthy individuals from 1KGP. To sum up, MIs of cancers and healthy individuals show different preferences across chromosomes. Besides, the

average MI number in each chromosome in cancer samples is extremely higher than that in healthy samples.

In addition to the comparison of MI distribution on chromosomes between cancer and healthy samples, we also included the comparison of MI distribution among chromosomes in six cancers. The MI distributions among 24 chromosomes in the six cancers including esophageal cancer, bladder cancer, hepatocellular carcinoma, lung cancer, prostate cancer, and pancreatic cancer are shown in Figure 3.3. Roughly, MI count correlated with chromosome length in each type of cancer.

The result in Figure 3.2A shows that Chromosome 6, 8, 13 are observed differently with other chromosomes in MIs of esophageal cancer. Specifically, Chromosome 8 has much fewer MIs than Chromosome 7 and even fewer MIs than Chromosome 9, but the Chromosome 8 have more base pairs than Chromosome 9. The situation is the same with Chromosome 13. However, the situation is different with Chromosome 6. Besides, there are several differences of the chromosomes which MIs prefer to be located in among the six cancers.

Figure 3.3B implies that MIs from bladder cancer samples are less likely to happen on Chromosome 8 and more likely to occur on Chromosome 6 and 11. In addition, the result shows that the distribution of MIs among chromosomes in bladder cancer is very similar with that in esophageal cancer shown in Figure 3.3A.

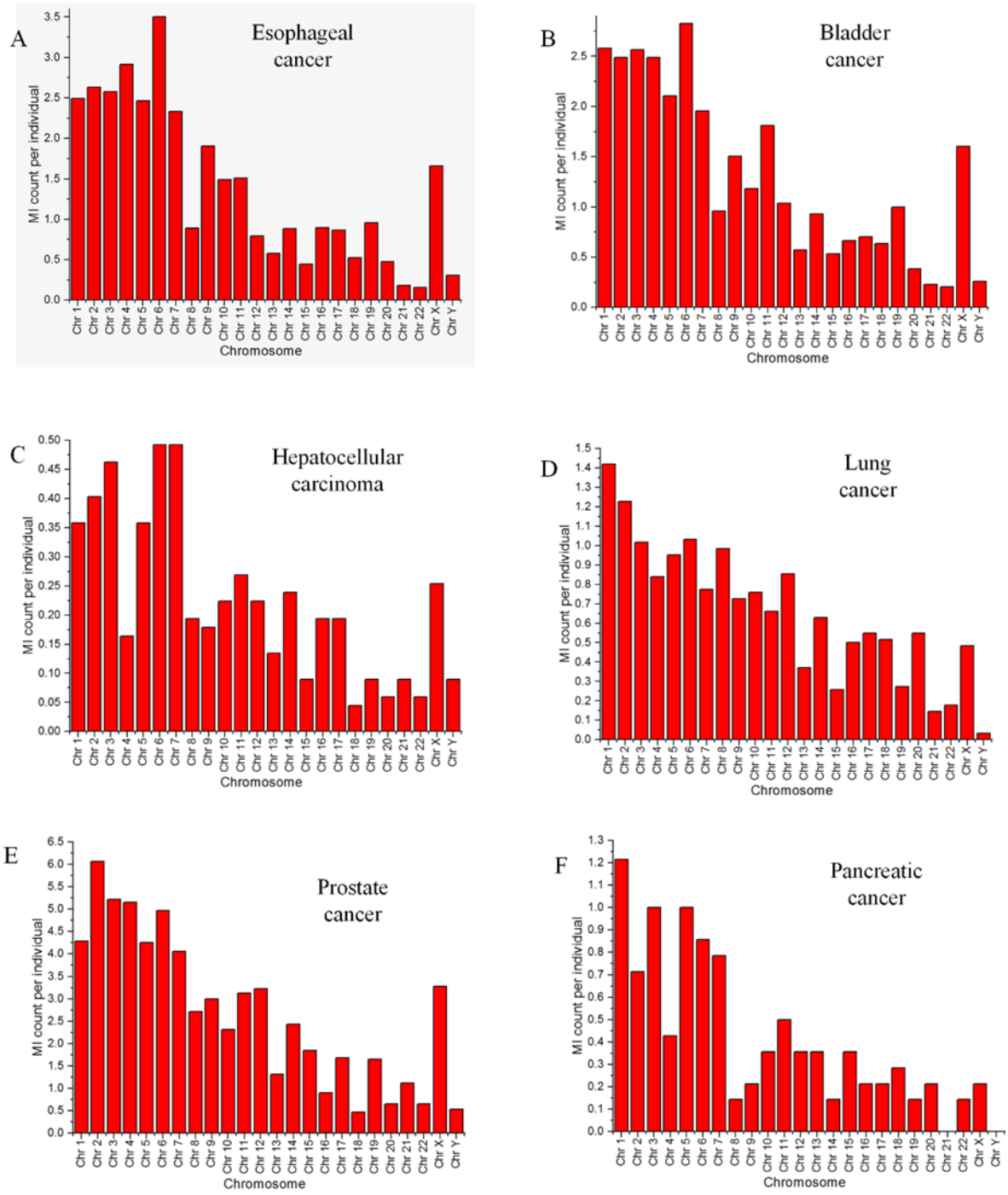

Figure 3.3 - MI distribution among 24 chromosomes in six cancers. (A) MI distribution among chromosomes in esophageal cancer. (B) MI distribution among chromosomes in bladder cancer. (C) MI distribution among chromosomes in hepatocellular carcinoma. (D) MI distribution among chromosomes in lung cancer.

(E) MI distribution among chromosomes in prostate cancer. (F) MI distribution among chromosomes in pancreatic cancer.

Figure 3.3C shows that MIs from hepatocellular carcinoma samples are more likely to happen on Chromosome 6 and 7 and less likely to happen on Chromosome 4. Figure 3.3D shows that MIs from lung cancer are more likely to happen on the Chromosome 6, 8, and 12. Figure 3.3E displays that MIs from prostate cancer are more possible to happen on the Chromosome 2, 6 and 14. Figure 3.3F implies that MIs from pancreatic cancers are more likely to happen on Chromosome 5 and 11, and less likely to happen on Chromosome 2, 4 and 8. It should be noted that, only lung cancer and pancreatic cancer show higher number of MIs in Chromosome 1 than Chromosome 2.

Besides, the MIs located on Chromosome 6 occurred more frequently in all the five cancers including esophageal cancer, bladder cancer, hepatocellular carcinoma, lung cancer, prostate cancer except only pancreatic cancer. In summary, MIs in the six cancers all show different prevalence across 24 chromosomes. The MI preferences for disparate chromosomes among the six cancers could provide a guidance for diagnosis and therapy on cancers in the future. Medical doctors should focus more on those chromosomes that MIs frequently occur.

3.3.2 Average number of MIs per individual

The error bar plot of the average number of MIs per individual in all the six cancers is displayed in Figure 3.4. In general, the average number of MIs per individual in healthy samples (1.89) from the 1KGP is much lower than that of any type of cancer (The lowest number of MIs per individual in six cancers is hepatocellular carcinoma with 5.36).

Additionally, there is a huge difference of the average number of MIs in different cancers. Specifically, the average number of MIs per individual varies from 5.36 to 21.98 in the six cancers. The average MI count in prostate cancer is the highest with 24.18, while the pancreatic cancer individuals have the lowest MI count per individual with 5.36. This analysis also displays that MIs occur more frequently in cancer samples than in healthy samples. Moreover, the MI average number varies greatly from cancer to cancer.

3.3.3 Venn diagram of genes overlapped with MIs

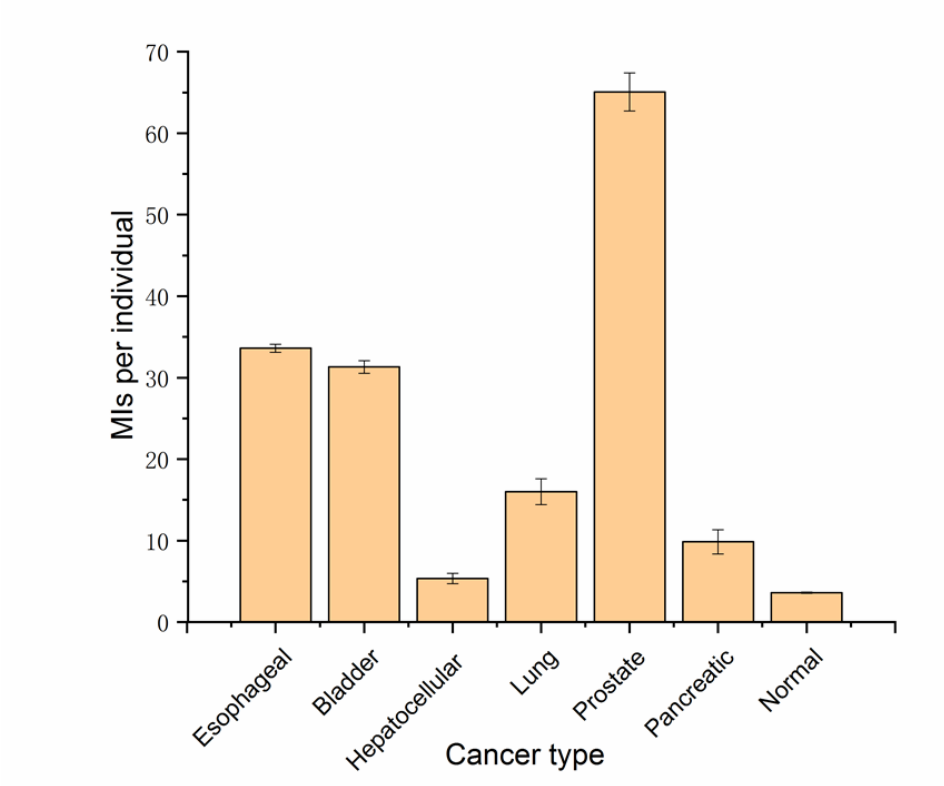

Figure 3.4 - Average number of MIs among individuals with six cancers and normal individuals.

To further interrogate the genes which are frequently found overlapped with MIs in the genomes of six cancers, we implemented a gene sharing analysis by the Venn diagram.

Since the Venn diagram is only available for at most five categories, we removed pancreatic cancer with the least samples of 14 and included the rest five cancers including esophageal cancer, bladder cancer, prostate cancer, lung cancer, and hepatocellular carcinoma in this analysis. Pointedly, we applied the unique genes overlapped with MIs of the five cancers into this analysis.

The Venn diagram of the shared genes by five cancers is shown in Figure 3.5. Generally, each cancer has more private genes of themselves than the ones shared by at least two cancers. Only three genes are shared in common by all the five cancers. The commonly shared three genes are AC079807.4, CNTNAP2, and EYS. AC079807.4. Interestingly, EYS gene was also found with MIs in healthy individual genomes from the 1KGP, while CNTAP2 gene was only discovered in cancer genomes but not healthy genomes. Moreover, the protein expression of CNTNAP2 was reported as disease-specific (DSS) [clxiv]. Besides, though CNTNAP2 gene have not been studied as traditional cancer mutational targets, CNTNAP2 was reported to action as tumor suppressors since they exhibit expression loss in multiple types of tumor [clxv]. Considering that some MIs in cancer individuals may also exist in healthy samples, we will target more on the MIs discovered in only cancer samples, but not found in the healthy individual genomes in the next part. Besides, we speculate that MIs in healthy samples will not be deleterious to health and will not directly cause cancers.

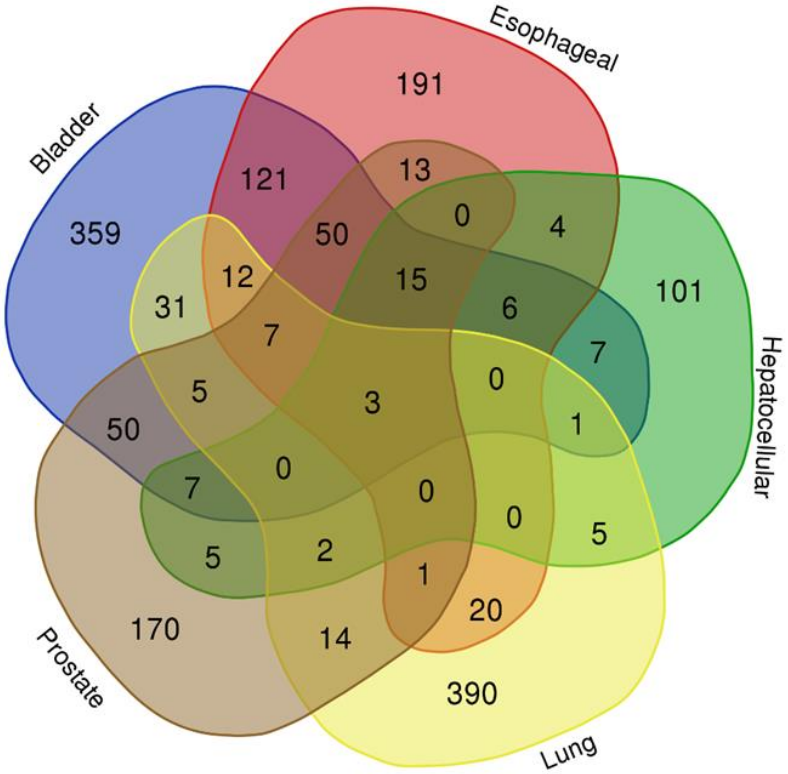

Figure 3.5 - Venn diagram of genes that overlap with MIs in individuals of five cancers. Esophageal, Bladder, Prostate, Hepatocellular, Lung, and Hepatocellular refers to esophageal cancer, bladder cancer, prostate cancer, lung cancer, and hepatocellular carcinoma individuals respectively.

We also performed cluster analysis with all the MIs of the six cancers (Figure 3.6). As it's shown in Figure 3.6, the six cancers were clustered into two branches, lung cancer and hepatocellular carcinoma into one branch and esophageal, bladder cancer, prostate cancer, and pancreatic cancer into another branch. Besides, the esophageal cancer and bladder cancer seem to be more closed in gene distance.

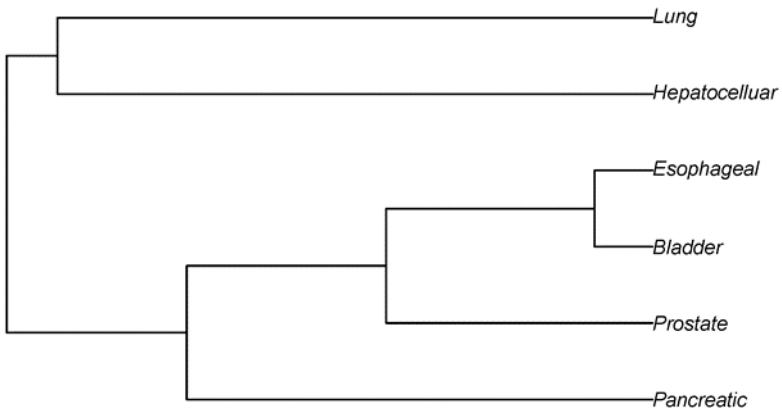

Figure 3.6 - Cluster result of the six cancers. Esophageal, Bladder, Prostate, Hepatocellular, Lung, Pancreatic, and Hepatocellular refers to esophageal cancer, bladder cancer, prostate cancer, lung cancer, pancreatic cancer, and hepatocellular carcinoma individuals respectively.

3.3.4 Top genes overlapped with MIs

We also summed up the top five genes in which MIs most frequently occurred among the healthy samples and the six cancers in Table 3.2 (A) and (B). We also listed the information if this gene with frequent MIs in cancer genomes also exist in healthy genomes. The top five genes that are most commonly presented with MIs in healthy samples are SLCA3A1, PREPL, SQSTM1, TNIK, and CTTNBP2. For all the six caners, the most frequent genes overlapped with MIs are SQSTM1, TNIK, CTTNBP2, NLRP4, RP11-380P13.1. Among them, SQSTM1, TNIK, and CTTNBP2 appeared in both healthy and cancer genomes. TNIK was reported to be related with distant recurrence in patients of stage II and III colorectal cancer [clxvi]. Also, SQSTM1 was recorded to be related with gastric cancer [clxvii]. Furthermore, there is evidence supporting that CTTNBP2 are potentially related to prostate cancer [clxviii].

Table 3.2 - Top five genes overlapped with MIs.

(A) Top five genes overlapped with MIs in normal samples.

Rank	Gene Name	Number of MIs overlap
1	SLC3A1	209
2	PREPL	209
3	SQSTM1	201
4	TNIK	184
5	CTTNBP2	136

(B) Top five genes overlapped with MIs in all six cancers.

Rank	Gene Name	Number of MIs overlap	If in healthy samples
1	SQSTM1	183	Yes
2	TNIK	137	Yes
3	CTTNBP2	132	Yes
4	NLRP4	124	No
5	RP11-380P13.1	112	Yes

(C) Top five genes overlapped with MIs in esophageal cancer.

Rank	Gene Name	Number of MIs overlap	If in healthy samples
1	SQSTM1	96	Yes
2	CTTNBP2	65	Yes
3	RERE	59	Yes
4	NLRP4	59	No
5	RP11-380P13.1	57	Yes

(D) Top five genes overlapped with MIs in bladder cancer.

Rank	Gene Name	Number of MIs overlap	If in healthy samples
1	TNIK	60	Yes
2	SQSTM1	59	Yes
3	CTTNBP2	46	Yes
4	RP11-380P13.1	44	Yes
5	RERE	42	Yes

(E) Top five genes overlapped with MIs in hepatocellular carcinoma.

Table 3.2 (continued)

Rank	Gene Name	Number of MIs overlap	If in healthy samples
1	TIAM2	9	No
2	LPP	7	No
3	AC002454.1	6	No
4	TULP4	4	Yes
5	AKNAD1	4	No

(F) Top five genes overlapped with MIs in lung cancer.

Rank	Gene Name	Number of MIs overlap	If in healthy samples
1	SORCS1	3	No
2	AL163953.3	3	No
3	RP11-420N3.2	3	No
4	MAP3K13	2	No
5	RPA3-AS1	2	No

(G) Top five genes overlapped with MIs in prostate cancer.

Rank	Gene Name	Number of MIs overlap	If in healthy samples
1	TNIK	23	Yes
2	NLRP4	22	No
3	SQSTM1	22	Yes
4	SLC3A1	21	Yes
5	FBXL7	20	Yes

(H) Top five genes overlapped with MIs in pancreatic cancer.

Rank	Gene Name	Number of MIs overlap	If in healthy samples
1	TNIK	4	Yes
2	SQSTM1	4	Yes
3	C12orf75	2	No
4	ROCK1	2	Yes
5	PRIM2	2	Yes

In summary, although the three genes all appeared with frequent MIs in both heathy genomes and cancer genomes, previous studies indicated that the three genes SQSTM1, TNIK, and CTTNBP2 displayed polymorphisms in different cancers and healthy individuals. Thus, these three genes should still be paid more attention to. It should be noted that, NLRP4 gene did not appear in the healthy individuals, but existed in cancer genomes. Not only that, NLPR4 was reported to be overexpressed in bladder carcinoma patients [clxix]. This result mean that the MIs in NLPR4 may have potential correlation with cancer occurrence.

We also placed the top five genes in each of the six cancers including hepatocellular carcinoma, lung cancer, pancreatic cancer, bladder cancer and pancreatic cancer in Table 3.2 (C) – (H). Generally speaking, the six cancers appeared different preferences for genes overlapped with frequent MIs. Besides the three genes of SQSTM1, TNIK, CTTNBP2 which were in the top five genes of healthy samples, also appeared in most cancers. The top genes overlapped with MIs in hepatocellular and lung cancer almost didn't appear in healthy individuals. This consequence may be due to the fact that not enough MIs overlapped with these genes compared with the number of MIs overlapped with genes in other cancers. More samples in the future may help solve this problem.

Interestingly, TIAM2 gene in the top five genes of hepatocellular carcinoma was reported by previous studies as overexpressed in over 86% of hepatocellular carcinoma patients [clxx]. Thus, high number of MIs in TIAM2 may further explain the relation of hepatocellular carcinoma and TIAM2 gene.

3.3.5 Comparison with SNPs

Considering the short length of MIs, we speculated that MI sites on the human genomes have been improperly reported by the previous studies as SNPs. To see the relationship of MIs and SNPs, we have downloaded the list of the exact sites of mutations in genes annotated in the Catalogue Of Somatic Mutations In Cancer (COSMIC) database, to see whether there is an overlap between MIs analyzed in this dissertation and the SNPs reported previously, 86,532 SNP coordinates were used to compare with the range of the unique 3,917MIs in the six cancers. A total of 139 SNPs in the coding regions were overlapped with MIs in this dissertation, which means that the 139 SNPs are possibly improperly reported by previous studies. Further studies need to re-examine whether the variations are SNPs or MIs. Table 3.3 list all the MIs that overlapped with the coding SNPs. The columns of Chr, MI start, MI end, Genes, C, and MI sequences represent the chromosomes that MIs and the overlapped SNPs located in, the overlapped MI start coordinates, MI end coordinates, the count of SNPs overlapped with MIs, and the MI sequence, respectively.

Plenty of the genes listed in Table 3.3 have been demonstrated to have association with cancers. For example, FBXO43 gene have been demonstrated to be differentially expressed in hepatitis B virus-related hepatocellular carcinoma [clxxi]. Besides, another gene PGF was found with increase in expression in individuals with severe colorectal cancer [clxxii]. The MIs that overlapped with these cancer-related genes frequently are interesting candidates for future studies.

3.4 Discussion

The comparison of MI counts in cancer samples and healthy individuals implies that the possibility of MI occurrence in cancer genomes (12,532 MIs/451 samples) is much higher than that in healthy samples (6,968 MIs/1937 samples). This result is understandable considering that the genomes of cancer samples are more flexible compared with the reference genome hg 19, which is assembled with a few healthy individuals. However, the frequency of MIs in gene regions of both the cancer samples (47.84%) and healthy samples (50.93%) is high, which implies the MIs in the regulatory regions may also have an influence on cancer genomes. Besides, MIs in cancer genomes tend to happen more frequently in exon regions than in intron regions compared with those in health individual genomes. This result also indicates that the MIs in cancer genomes are more devastating than those in healthy people. The MIs in intron regions may also be related with carcinogenesis mechanism. Reports indicate that a number of regulators, such as initiators, terminators, repressors and enhancers, as well as cancer-related SVs, occur in the intron regions [clxxiii]. In addition, some SVs such as MIs are usually extended in the reverse complementary regions. In fact, Scacheri et al summarized the variations in non-coding regions and found that SVs in non-coding regions of the human genomes also play important roles in cancers [clxxiv].

Our results show that MI distributions among chromosomes in six cancers have both similarities and differences. Some cancers seem to share some common chromosomes. For example, the MIs are more likely to happen on Chromosome 6 in almost all the six cancer genomes. The MI preferences for divergent chromosomes among six cancers may provide a guidance for the remedy and diagnosis on the five cancers. More attention should be paid on the chromosomes that MIs frequently occur in the future.

The comparing analysis of the top genes overlapped with the frequently occurring MIs between healthy individual genomes and cancer genomes is important for the future studies of comparative genome analysis between cancer genomes and healthy genomes. Interestingly, NLRP4 gene was not in genes of healthy samples, but it did appear in the cancer genomes. Not only that, NLPR4 was reported to be overexpressed in bladder carcinoma patients. This result mean that the MIs in NLPR4 may have potential correlation with cancer occurrence.

Generally speaking, the six cancers appeared different preferences for genes overlapped with frequent MIs. Besides the three genes of SQSTM1, TNIK, CTTNBP2 which appeared in most cancers, each cancer has its own private preferred genes with MIs. The top genes overlapped with MIs in hepatocellular and lung cancer almost didn't appear in healthy individuals. This consequence may be due to the fact that not enough MIs overlapped with these genes compared with the number of MIs overlapped with genes in other cancers. Interestingly, the TIAM2 gene in hepatocellular carcinoma was reported by previous studies as overexpressed in over 86% of hepatocellular carcinoma patients [170].Thus, the high number of MIs in TIAM2 may further explain the relation of hepatocellular carcinoma and TIAM2 gene.

The comparison of the MIs with the corresponding SNPs show that that the locations of MIs on the human genomes detected in this dissertation may have been improperly reported by the previous studies as SNPs. Moreover, the cancer landscape reflected by only SNP data may not be complete, and our MI data may also be a good supplement to the current studies on the relationship of genome variation and cancers.

Table 3.3 - The list of MIs overlapped with SNPs.

Chr	MI start	MI end	Gene	C	MI sequence
1	16,255,341	16,255,374	SPEN	4	AAAACCACGCAAGGCATGAGTTCCAGGCGGTTTT
1	231,339,704	231,339,728	TRIM67	3	CCCCAGGATGTAGCCATCCACGGGG
5	149,631,586	149,631,612	CAMK2A	2	TGGATATCCCTCCCCAGAAGTGCTGCG
5	156,741,383	156,741,417	CYFIP2	3	TCAGACGAAGAGCTCGCGATACTCCTCGTCTGAC
2	149,851,012	149,851,030	KIF5C	4	CTCTTCGCTGAGCGAGTCC
2	179,499,110	179,499,138	TTN	16	GCTGAGGTGGAGGGCAAGAAGACCTCAGC
3	182,871,613	182,871,642	LAMP3	3	CAGCTGCCCACAATACCACCCGCACAGCTG
4	119,951,540	119,951,558	SYNPO2	6	GACTCTTCCTCCTTTCTTT
4	155,252,824	155,252,851	DCHS2	7	TCTATGTCATAACCTGGGCAGATGGTGC
6	25,973,440	25,973,473	TRIM38	2	CTTGGAAACATTGCACATAGTATGAAGTTCCAAG
8	101,154,370	101,154,403	FBXO43	10	TTTGCAGAAACAGAGATTTTGAAGATGTCGCAAA
9	91,628,348	91,628,369	SHC3	2	AGCAGTGACCTGGCCAGCACTG
9	95,782,696	95,782,727	FGD3	2	CATCCCGGGCCACTGACTCACAGGCATGTCTG
11	2,439,570	2,439,599	TRPM5	4	CCCCGTCTCAGAGGATCTCCAGGGCCGTGG
11	104,815,538	104,815,565	CASP4	4	TTGAAACTCCAAGGGCCAAAGCTCAAAT
11	129,772,306	129,772,340	PRDM10	5	CCAGGTCGAGCCACCTGCACACAGTGACTCCCTG
11	134,051,038	134,051,068	NCAPD3	4	GCAGGTGTGTGGGGATGTACTCTCCACCTGC
12	76,739,741	76,739,773	BBS10	3	ACTGCTTACCAATCAACCCTTGGTAAGCAGTCA
14	75,416,174	75,416,204	PGF	12	GCTGGTGGACGTCGTGTCCGAGTACCCCAGC
15	26,793,120	26,793,147	GABRB3	20	GCCCATGCCTCGAGAAGGGCATGGGCGA
16	14,697,990	14,698,016	PARN	2	ATGCCAAAGAACAGGTAATTGGGCTAT
16	70551522	70551541	COG4	4	TTCCCTGGCCAAGGTAAGGC
17	45,896,450	45,896,477	OSBPL7	1	CTTCACCCCCAGATCACCAAGGGGAAGC
17	77,758,078	77,758,099	CBX2	5	AGGCCGCACTTGTTGGTGGCCT
19	38,376,145	38,376,176	WDR87	4	TTTTTCCAAGTGCCCATGCCTCTGTGGAAAAA
20	62,738,141	62,738,163	NPBWR2	2	GAGCCCCTTGACAGCAGGGGCTC
X	7,889,850	7,889,872	PNPLA4	3	TTCTTCCTCCCAGCGCTCACGAG
X	70,316,615	70,316,635	FOXO4	2	CCCCAGGATTCCGGGCTGGGG

The columns of Chr, MI start, MI end, Genes, C, and MI sequences represent the chromosomes that MIs and the overlapped SNPs located in, the overlapped MI start coordinates, MI end coordinates, the count of SNPs overlapped with MIs, and the MI sequence, respectively.

3.5 Conclusions

In this chapter, we have generated a comprehensive analysis of MIs of 451 samples of six cancers, including 145 samples of esophageal cancer, 131 samples of bladder cancer, 67 samples of hepatocellular carcinoma, 62 samples of lung cancer, 32 samples of prostate cancer, and 14 samples of pancreatic cancer from SRA database. We also used the 1,937 healthy individuals from the 1KGP as the control samples. In all the cancer samples, we detected a total of 12,532 MIs, of which 5,995 (47.84%) overlapped with gene regions. Of the total 12,532 MIs, 3,917 MIs are unique. In other words, the rest 8,615 MIs are repetitive to at least one MI in the 12,532 MIs.

We further analyzed the distribution of MIs in the six cancers among 24 chromosomes. The MI distributions among chromosomes among different cancers have both similarities and differences. Additionally, we analyzed the average MI count per individual among each cancer and found that prostate cancer had the most MIs with about average 65 MIs per genome. In addition, we studied the top genes that overlapped with frequent MIs in six cancers and healthy samples. The results showed that the genes where MIs frequently occurred in different cancers were distinct. The MI preferences for divergent genes among six cancers could provide a navigation for the surgery and diagnosis of the six cancers. Besides, through the contrast of MIs detected in this dissertation and the SNPs in coding regions reported in the previous studies, an overlap was found between MIs and SNPs. Our MI data will be a good supplement to the current studies on the relationship of genome variation and cancers.

All these results above indicate that MIs have a potential association with cancer development. The comparative analysis of MIs in the six types of cancer will help further understand precision medicine on cancer individuals. At the same time, we hope our MI analysis will help future studies discover the disease mechanisms.

One limitation of this study is the small sample size and the number of MIs. Due to the fact that MI count in each sample is not as high as that of SNPs, 12,532 MIs from 451 samples were detected. As a next plan, we intend to include samples from more cancers and obtain more MIs to broaden the research cohort. Besides, we will focus more on the MIs that only exist in only one specific cancer. Thus, we will better explore the relationship of MIs and a specific cancer.

CHAPTER 4 CONCLUSIONS AND FUTURE DIRECTIONS

The advent of high-throughput sequencing have brought abundant sequencing data in recent years. The large amount of personal genome sequencing databases such as 1KGP and SRA have provided solid foundation for the SV analysis. At the same time, plenty of short reads generated by high-throughput sequencing have brought a series of problems and challenges. The current SV databases have the problem of incomplete and inconsistent SV information since the SVs are obtained through different tools and experiments. Accounting that there is currently no general guideline or tool for evaluating these SV sites, these SV-site information may not be absolutely reliable to a large extent. Besides, the studies about the analysis of the common or overlapped SVs among different genomes are lacked. This indicates that people's understanding of genomic SVs is till limited. Undoubtedly, future researches are needed to yield a more accurate and reliable database of such SV analysis among different genomes. To achieve this goal and to enrich disease-related genomic structural variation, future researches need to focus on how to construct a credible, accurate, genome-level overview of the human genome SVs.

In this dissertation, MIs are defined as inversions that are shorter than 100 bp and larger than 10 bp. MIs are an important type of SVs, which could have significant influence on multiple diseases, population diversity and human evolution. In fact, although the researches and understanding of genomic SVs are developing rapidly, there are still considerable limitations of the field of MI studies. Thus, we made a comprehensive

analysis of the small-scale MIs in both healthy genomes and cancer genomes with a series of bioinformatics methods. For the analysis of MIs in healthy genomes, we made a landscape of MIs on non-disease individuals from the 1KGP based on high-throughput sequencing data. Besides, we included the MIs detected in seven non-human primate genomes including chimpanzee, gorilla, orangutan, gibbon, baboon, rhesus monkey, and squirrel monkey from UCSC Genome Browser to make a comparative genome analysis. With the software MID for 26 human populations and searchUMI for alignments between the human and seven non-human primate assembles, we obtained a total of 6,968 MIs in 1,937 individuals of the 26 populations from the 1KGP and 24,476 MIs in the seven non-human primate genomes.

The MI results showed that the MIs from five super-populations (Africa, America, Europe, South Asia, and East Asia) revealed population structures at divergent levels and may affect individual genomes in several aspects. Specifically, the average number of MIs per genome in the five super-populations was in linear relationship with the descending order: Africa > America > Europe > South Asia > East Asia and this descending order also coincided with the “Out of Africa” hypothesis. We inferred that after the Out-of-Africa migration of modern human ancestors and hybridization among the four non-African super-populations, the ancient Africans introduced some MIs into the ancestors of the other four super-populations. Besides, the Africans had the most MIs shared with the seven non-human primates among the five super-populations. The phylogenetic tree and PCA results of MIs showed that the widespread MIs in human genomes were related with geographical distribution and human migration. In general, our results of plenty of MIs are good supplements to existing evolution markers in reconstructing human evolutionary history

and the MIs in particular super-populations, especially those overlapped with gene regions, are informative for understanding the association of health and genetic variations.

To further explore the roles of MIs in cancers, we detected and analyzed the MIs of 451 samples based on the high-throughput sequencing data of six cancers, including esophageal cancer, hepatocellular carcinoma, lung cancer, pancreatic cancer, prostate cancer and bladder cancer from SRA database. We also applied the 1,937 samples from the 1KGP analysis as the control samples. The MI distribution among chromosomes in different cancers have both similarities and differences. When we counted the average MI number per individual, we found that the MI number in cancer genomes was much higher than that in healthy genomes. Besides, the genes in which MIs frequently occurred were distinct in six cancers. Besides, an overlap was found between the MIs detected in this dissertation and the SNPs reported in previous studies.

Generally, our analysis of MIs in the six cancer genomes have shown that the amount of MIs in different cancers and the preference for chromosomes and genes are all different. All the results mentioned above indicate that MIs have a potential association with cancer development. We believe that our MI data will be a good supplement to the current studies on the relationship of genome variation and cancers. The comparative analysis of MIs in the six types of cancer may benefit for further employment of precision medicine on cancer individuals. In addition, our MI analysis may help future studies discover the disease mechanisms. In future, we plan to include more kinds of cancers and detect more MIs to enlarge the cohort of this study. Besides, we intend to include the analysis of the location of MIs on the chromosomes. For example, we could explore whether MIs tend to be near

telomere, centromeres and other special locations. Besides, it will also be significant to explore the GC content and the expression level of genes near the MIs. We believe these analyses will help reveal the mechanism of MIs. In addition, different cancers may have different preference for MI length distribution. We infer that whether MIs with different length are inherited or eliminated by evolution depends on different evolution pressure. Thus, we believed that these analysis will benefit analyzing the relationship of MIs, molecular evolution and genetic stability.

In general, this dissertation has given a comprehensive analysis of the long-neglected MIs based on high-throughput sequencing data with a series of bioinformatics methods. The MI analysis results indicate that MIs have an undeniable significance on human diversity, evolution, environmental adaption, and diseases. We hope our analysis of MIs could improve our understanding of SVs in human diversity and evolution. Besides, we expect the comparative analysis of MIs in different cancer genomes could provide further insight into precision medicine and disease mechanisms.

APPENDIX A. THE LIST OF SAMPLES FROM 1KGP AND SRA

The list of 1,932 samples from 1KGP classified by super-populations and populations

East Asia (CHB, JPT, CHS, CDX, KHV):

CHB includes 73 samples:

NA18525, NA18526, NA18528, NA18530, NA18531, NA18532, NA18533, NA18534, NA18535, NA18536, NA18537, NA18538, NA18539, NA18541, NA18542, NA18543, NA18544, NA18545, NA18546, NA18547, NA18553, NA18560, NA18565, NA18567, NA18572, NA18574, NA18579, NA18591, NA18593, NA18596, NA18599, NA18602, NA18603, NA18605, NA18606, NA18608, NA18609, NA18611, NA18612, NA18613, NA18615, NA18616, NA18617, NA18618, NA18619, NA18620, NA18621, NA18622, NA18623, NA18624, NA18625, NA18626, NA18627, NA18628, NA18629, NA18630, NA18631, NA18632, NA18633, NA18634, NA18635, NA18636, NA18637, NA18639, NA18641, NA18642, NA18643, NA18644, NA18646, NA18647, NA18648, NA18748, NA18749.

JPT includes 90 samples:

NA18939, NA18942, NA18943, NA18944, NA18945, NA18946, NA18947, NA18948, NA18949, NA18950, NA18951, NA18953, NA18954, NA18956, NA18957, NA18959, NA18960, NA18961, NA18963, NA18964, NA18965, NA18966, NA18967, NA18968,

NA18970, NA18971, NA18972, NA18973, NA18974, NA18975, NA18976, NA18978, NA18979, NA18980, NA18981, NA18982, NA18983, NA18984, NA18986, NA18987, NA18988, NA18989, NA18990, NA18991, NA18993, NA18994, NA18995, NA18998, NA18999, NA19000, NA19001, NA19003, NA19004, NA19005, NA19006, NA19007, NA19009, NA19010, NA19011, NA19012, NA19054, NA19055, NA19056, NA19057, NA19058, NA19059, NA19060, NA19062, NA19063, NA19064, NA19065, NA19066, NA19067, NA19068, NA19070, NA19072, NA19074, NA19076, NA19077, NA19078, NA19079, NA19080, NA19082, NA19083, NA19085, NA19086, NA19087, NA19088, NA19090, NA19091.

CHS includes 80 samples:

HG00403, HG00404, HG00406, HG00407, HG00409, HG00410, HG00419, HG00421, HG00422, HG00428, HG00436, HG00437, HG00442, HG00443, HG00445, HG00446, HG00448, HG00449, HG00451, HG00452, HG00457, HG00458, HG00463, HG00464, HG00472, HG00473, HG00478, HG00479, HG00501, HG00525, HG00533, HG00534, HG00536, HG00537, HG00542, HG00543, HG00556, HG00557, HG00559, HG00560, HG00566, HG00577, HG00578, HG00580, HG00581, HG00583, HG00592, HG00593, HG00595, HG00598, HG00599, HG00611, HG00620, HG00622, HG00623, HG00625, HG00626, HG00628, HG00629, HG00631, HG00632, HG00634, HG00656, HG00657, HG00662, HG00663, HG00671, HG00674, HG00675, HG00683, HG00684, HG00692, HG00693, HG00699, HG00702, HG00705, HG00707, HG00708 , HG00717, HG00728.

CDX includes 62 samples:

HG00844, HG00864, HG00867, HG00879, HG00956, HG00978, HG00982, HG01029, HG01031, HG01794, HG01795, HG01796, HG01797, HG01798, HG01799, HG01801, HG01802, HG01805, HG01806, HG01807, HG01810, HG01811, HG01812, HG01813, HG01815, HG02151, HG02152, HG02154, HG02164, HG02166, HG02178, HG02180, HG02181, HG02182, HG02184, HG02185, HG02186, HG02188, HG02190, HG02356, HG02367, HG02371, HG02373, HG02374, HG02375, HG02377, HG02380, HG02382, HG02386, HG02387, HG02388, HG02390, HG02392, HG02394, HG02395, HG02397, HG02401, HG02402, HG02406, HG02407, HG02408, HG02410.

KHV includes 74 samples:

HG01595, HG01597, HG01598, HG01599, HG01840, HG01842, HG01843, HG01844, HG01845, HG01846, HG01849, HG01850, HG01852, HG01855, HG01860, HG01861, HG01862, HG01863, HG01864, HG01865, HG01867, HG01868, HG01869, HG01870, HG01871, HG01872, HG01873, HG01874, HG01878, HG02016, HG02017, HG02019, HG02024, HG02025, HG02026, HG02028, HG02029, HG02035, HG02040, HG02046, HG02048, HG02057, HG02058, HG02060, HG02061, HG02064, HG02067, HG02069, HG02070, HG02072, HG02073, HG02075, HG02076, HG02079, HG02081, HG02082, HG02086, HG02113, HG02116, HG02121, HG02122, HG02127, HG02128, HG02133, HG02134, HG02136, HG02137, HG02138, HG02139, HG02142, HG02512, HG02513

HG02521, HG02522.

Europe (CEU, TSI, FIN, GBR, IBS):

CEU includes 80 samples:

NA06984, NA06985, NA06986, NA06989, NA07000, NA07037, NA07048, NA07056, NA07346, NA07347, NA10847, NA11829, NA11831, NA11832, NA11881, NA11892, NA11893, NA11894, NA11918, NA11919, NA11920, NA11930, NA11931, NA11932, NA11933, NA11992, NA11994, NA12004, NA12005, NA12006, NA12044, NA12045, NA12046, NA12058, NA12144, NA12154, NA12155, NA12156, NA12234, NA12249, NA12272, NA12273, NA12275, NA12282, NA12283, NA12286, NA12340, NA12341, NA12342, NA12347, NA12348, NA12383, NA12399, NA12400, NA12413, NA12414, NA12489, NA12546, NA12716, NA12717, NA12718, NA12748, NA12749, NA12760, NA12762, NA12775, NA12776, NA12777, NA12778, NA12812, NA12814, NA12827, NA12829, NA12830, NA12842, NA12843, NA12872, NA12874, NA12889, NA12890.

TSI includes 75 samples:

NA20503, NA20504, NA20505, NA20506, NA20507, NA20508, NA20509, NA20510, NA20511, NA20512, NA20513, NA20514, NA20515, NA20516, NA20517, NA20520, NA20521, NA20522, NA20525, NA20526, NA20527, NA20528, NA20529, NA20530, NA20531, NA20532, NA20533, NA20535, NA20538, NA20539, NA20541, NA20543,

NA20544, NA20587, NA20588, NA20589, NA20752, NA20753, NA20754, NA20755, NA20756, NA20757, NA20759, NA20761, NA20762, NA20764, NA20766, NA20767, NA20768, NA20772, NA20775, NA20778, NA20783, NA20785, NA20786, NA20787, NA20790, NA20795, NA20797, NA20801, NA20803, NA20805, NA20806, NA20807, NA20809, NA20810, NA20813, NA20818, NA20819, NA20821, NA20822, NA20826, NA20827, NA20828, NA20832.

FIN includes 77 samples:

HG00171, HG00173, HG00174, HG00176, HG00177, HG00178, HG00179, HG00180, HG00181, HG00183, HG00185, HG00186, HG00188, HG00189, HG00190, HG00266, HG00267, HG00268, HG00269, HG00271, HG00272, HG00273, HG00274, HG00275, HG00276, HG00277, HG00278, HG00280, HG00281, HG00284, HG00285, HG00288, HG00290, HG00304, HG00306, HG00308, HG00309, HG00310, HG00311, HG00313, HG00319, HG00320, HG00324, HG00325, HG00326, HG00327, HG00329, HG00330, HG00332, HG00335, HG00336, HG00337, HG00341, HG00344, HG00345, HG00349, HG00350, HG00351, HG00353, HG00355, HG00356, HG00357, HG00358, HG00360, HG00361, HG00362, HG00364, HG00365, HG00367, HG00368, HG00371, HG00372, HG00378, HG00379, HG00382, HG00383, HG00384.

GBR includes 74 samples:

HG00096, HG00099, HG00100, HG00101, HG00102, HG00103, HG00105, HG00106, HG00107, HG00110, HG00111, HG00112, HG00113, HG00114, HG00115, HG00116, HG00117, HG00118, HG00119, HG00120, HG00121, HG00122, HG00123, HG00124, HG00126, HG00127, HG00129, HG00131, HG00132, HG00133, HG00136, HG00137, HG00138, HG00139, HG00140, HG00141, HG00142, HG00143, HG00148, HG00149, HG00151, HG00154, HG00157, HG00160, HG00231, HG00233, HG00235, HG00236, HG00238, HG00239, HG00240, HG00242, HG00243, HG00244, HG00245, HG00246, HG00250, HG00251, HG00252, HG00253, HG00254, HG00255, HG00256, HG00258, HG00259, HG00260, HG00261, HG00262, HG00263, HG00264, HG00265, HG01334, HG01790, HG01791.

IBS includes 70 samples:

HG01501, HG01503, HG01506, HG01507, HG01509, HG01510, HG01512, HG01513, HG01515, HG01516, HG01518, HG01519, HG01521, HG01522, HG01524, HG01525, HG01527, HG01528, HG01530, HG01531, HG01536, HG01537, HG01606, HG01607, HG01607, HG01610, HG01619, HG01628, HG01630, HG01631, HG01632, HG01669, HG01670, HG01672, HG01676, HG01680, HG01682, HG01684, HG01685, HG01694, HG01695, HG01704, HG01705, HG01707, HG01708, HG01709, HG01710, HG01746, HG01747, HG01756, HG01757, HG01761, HG01762, HG01765, HG01766, HG01767, HG01768, HG01770, HG01775, HG01777, HG01779, HG01781, HG01784, HG01785, HG01786, HG02219, HG02221, HG02224, HG02232, HG02233.

Africa (YRI, LWK, GWD, MSL, ESN, ASW, ACB):

YRI includes 50 samples:

NA18489, NA18499, NA18501, NA18502, NA18505, NA18508, NA18510, NA18517, NA18519, NA18858, NA18861, NA18870, NA18876, NA18877, NA18878, NA18879, NA18881, NA19093, NA19095, NA19096, NA19098, NA19099, NA19102, NA19108, NA19113, NA19117, NA19121, NA19137NA19138, NA19144, NA19146, NA19147, NA19149, NA19153, NA19159, NA19175, NA19184, NA19189, NA19201, NA19206, NA19207, NA19209, NA19210, NA19214, NA19222, NA19225, NA19238, NA19239

NA19240, NA19257.

LWK includes 83 samples:

NA19017, NA19019, NA19023, NA19024, NA19025, NA19026, NA19028, NA19030, NA19031, NA19035, NA19036, NA19037, NA19038, NA19041, NA19042, NA19043, NA19307, NA19308, NA19310, NA19311, NA19313, NA19315, NA19316, NA19317, NA19318, NA19319, NA19320, NA19321, NA19323, NA19324, NA19327, NA19328, NA19331, NA19332, NA19334, NA19338, NA19346, NA19347, NA19350, NA19351, NA19355, NA19372, NA19374, NA19375, NA19377, NA19378, NA19385, NA19390, NA19391, NA19394, NA19397, NA19401, NA19403, NA19404, NA19429, NA19435, NA19437, NA19438, NA19439, NA19440, NA19443, NA19444, NA19445, NA19446,

NA19449, NA19451, NA19452, NA19453, NA19454, NA19455, NA19456, NA19461, NA19462, NA19463, NA19466, NA19467, NA19469, NA19470, NA19471, NA19472, NA19473, NA19474, NA19475.

GWD includes 102 samples:

HG02461, HG02462, HG02464, HG02465, HG02561, HG02562, HG02568, HG02570, HG02571, HG02573, HG02574, HG02582, HG02583, HG02585, HG02586, HG02588, HG02589, HG02594, HG02595, HG02610, HG02611, HG02613, HG02614, HG02620, HG02621, HG02623, HG02624, HG02628, HG02629, HG02634, HG02635, HG02642, HG02643, HG02645, HG02646, HG02666, HG02667, HG02675, HG02676, HG02678, HG02679, HG02702, HG02703, HG02715, HG02716, HG02721, HG02722, HG02756, HG02757, HG02760, HG02763, HG02768, HG02769, HG02771, HG02772, HG02798, HG02799, HG02804, HG02805, HG02807, HG02808, HG02810, HG02811, HG02813, HG02814, HG02816, HG02820, HG02837, HG02840, HG02851, HG02852, HG02860, HG02861, HG02870, HG02878, HG02879, HG02881, HG02882, HG02884, HG02885, HG02887, HG02888, HG02890, HG02891, HG02895, HG02896, HG03024, HG03027, HG03028, HG03039, HG03040, HG03045, HG03046, HG03048, HG03049, HG03240, HG03241, HG03246, HG03247, HG03258, HG03259, HG03539.

MSL includes 81 samples:

HG03052, HG03054, HG03055, HG03057, HG03058, HG03060, HG03061, HG03063, HG03066, HG03069, HG03072, HG03073, HG03074, HG03078, HG03079, HG03081, HG03082, HG03084, HG03085, HG03088, HG03091, HG03095, HG03096, HG03097, HG03209, HG03212, HG03224, HG03225, HG03376, HG03378, HG03380, HG03382, HG03385, HG03388, HG03391, HG03394, HG03397, HG03401, HG03410, HG03419, HG03428, HG03432, HG03433, HG03436, HG03437, HG03439, HG03442, HG03445, HG03446, HG03449, HG03451, HG03452, HG03455, HG03457, HG03458, HG03460, HG03461, HG03464, HG03469, HG03470, HG03472, HG03473, HG03476, HG03478, HG03479, HG03484, HG03485, HG03547, HG03548, HG03556, HG03557, HG03558, HG03559, HG03563, HG03567, HG03571, HG03572, HG03575, HG03577, HG03578, HG03583.

ESN includes 87 samples:

HG02922, HG02923, HG02944, HG02946, HG02947, HG02952, HG02968, HG02970, HG02973, HG02974, HG02976, HG02977, HG02979, HG02981, HG03099, HG03100, HG03103, HG03105, HG03108, HG03109, HG03111, HG03112, HG03114, HG03115, HG03117, HG03118, HG03120, HG03121, HG03123, HG03124, HG03126, HG03127, HG03129, HG03130, HG03130, HG03133, HG03135, HG03136, HG03139, HG03159, HG03160, HG03162, HG03163, HG03166, HG03168, HG03169, HG03172, HG03175, HG03189, HG03190, HG03195, HG03196, HG03198, HG03265, HG03267, HG03270, HG03271, HG03279, HG03280, HG03291, HG03294, HG03295, HG03297, HG03298, HG03301, HG03303, HG03304, HG03311, HG03313, HG03342, HG03343, HG0335,

HG03352, HG03354, HG03363, HG03367, HG03369, HG03370, HG03372, HG03499, HG03511, HG03514, HG03515, HG03517, HG03518, HG03520, HG03521.

ASW includes 58 samples:

NA19625, NA19700, NA19701, NA19703, NA19704, NA19707, NA19711, NA19712, NA19713, NA19818, NA19819, NA19834, NA19835, NA19900, NA19901, NA19904, NA19908, NA19909, NA19914, NA19916, NA19917, NA19920, NA19921, NA19923, NA19982, NA19984, NA19985, NA20126, NA20127, NA20276, NA20278, NA20281, NA20282, NA20287, NA20289, NA20291, NA20294, NA20296, NA20299, NA20314, NA20317, NA20318, NA20322, NA20332.

ACB includes 81 samples:

HG01879, HG01880, HG01882, HG01883, HG01885, HG01886, HG01889, HG01890, HG01894, HG01896, HG01912, HG01914, HG01915, HG01985, HG01986, HG01989, HG01990, HG02009, HG02010, HG02012, HG02014, HG02051, HG02052, HG02053, HG02095, HG02107, HG02108, HG02111, HG02143, HG02144, HG02255, HG02282, HG02283, HG02284, HG02307, HG02308, HG02309, HG02314, HG02315, HG02317, HG02318, HG02322, HG02323, HG02325, HG02330, HG02332, HG02339, HG02419, HG02427, HG02429, HG02433, HG02439, HG02442, HG02445, HG02449, HG02450, HG02455, HG02470, HG02471, HG02476, HG02477, HG02479, HG02481, HG02484,

HG02485, HG02489, HG02497, HG02501, HG02502, HG02505, HG02511, HG02536, HG02537, HG02541, HG02545, HG02546, HG02549, HG02554, HG02557, HG02558, HG02580.

America (MXL, PUR, CLM, PEL)

MXL includes 55 samples:

NA19648, NA19649, NA19651, NA19652, NA19654, NA19657, NA19661, NA19663, NA19669, NA19676, NA19678, NA19679, NA19682, NA19684, NA19685, NA19719, NA19720, NA19722, NA19723, NA19725, NA19726, NA19728, NA19729, NA19731, NA19732, NA19734, NA19735, NA19741, NA19746, NA19747, NA19749, NA19750, NA19752, NA19755, NA19756, NA19758, NA19759, NA19761, NA19762, NA19764, NA19770l, NA19773, NA19774, NA19776, NA19777, NA19779, NA19780, NA19782, NA19783, NA19785, NA19786, NA19788, NA19789, NA19792, NA19795.

PUR includes 39 samples:

HG00731, HG00732, HG00733, HG00742, HG00743, HG01048, HG01058, HG01063, HG01064, HG01077, HG01085, HG01086, HG01088, HG01089, HG01092, HG01095, HG01104, HG01161, HG01162, HG01164, HG01188, HG01200, HG01205, HG01286, HG01303, HG01305, HG01308, HG01311, HG01312, HG01325, HG01393, HG01395, HG01396, HG01398, HG01402, HG01405, HG01412, HG01413, HG01414.

CLM includes 70 samples:

HG01112, HG01113, HG01119, HG01121, HG01122, HG01124, HG01125, HG01133, HG01134, HG01136, HG01137, HG01140, HG01142, HG01149, HG01250, HG01251, HG01253, HG01254, HG01256, HG01257, HG01259, HG01260, HG01269, HG01271, HG01272, HG01277, HG01280, HG01284, HG01341, HG01342, HG01344, HG01345, HG01348, HG01350, HG01351, HG01353, HG01360, HG01362, HG01363, HG01366, HG01369, HG01372, HG01375, HG01378, HG01384, HG01431, HG01432, HG01435, HG01438, HG01441, HG01443, HG01444, HG01455, HG01461, HG01462, HG01464, HG01468, HG01474, HG01479, HG01485, HG01486, HG01488, HG01491, HG01494, HG01495, HG01497, HG01498, HG01550, HG01551, HG01556.

PEL includes 63 samples:

HG01565, HG01577, HG01578, HG01917, HG01918, HG01920, HG01923, HG01926, HG01927, HG01932, HG01933, HG01935, HG01938, HG01939, HG01941, HG01942, HG01944, HG01945, HG01947, HG01948, HG01951, HG01953, HG01954, HG01967, HG01968, HG01970, HG01971, HG01973, HG01974, HG01976, HG01977, HG01979, HG01982, HG01992, HG01997, HG02002, HG02003, HG02008, HG02089, HG02090, HG02104, HG02105, HG02146, HG02150, HG02252, HG02260, HG02265, HG02271, HG02274, HG02275, HG02277, HG02278, HG02285, HG02291, HG02292, HG02298, HG02299, HG02301, HG02304, HG02312, HG02345, HG02348, HG02425.

South Asia (GIH, PJL, BEB, STU, ITU):

GIH includes 83 samples:

NA20845, NA20846, NA20847, NA20849, NA20850, NA20852, NA20853, NA20858, NA20859, NA20862, NA20863, NA20864, NA20867, NA20868, NA20869, NA20870, NA20871, NA20872, NA20877, NA20878, NA20881, NA20882, NA20884, NA20885, NA20886, NA20887, NA20888, NA20889, NA20890, NA20892, NA20893, NA20894, NA20895, NA20896, NA20897, NA20898, NA20899, NA20900, NA20901, NA20902, NA20904, NA20905, NA20906, NA20908, NA20910, NA20911, NA21086, NA21087, NA21088, NA21089, NA21090, NA21091, NA21092, NA21093, NA21094, NA21095, NA21098, NA21099, NA21100, NA21102, NA21103, NA21104, NA21105, NA21106, NA21107, NA21108, NA21110, NA21114, NA21115, NA21118, NA21119, NA21124, NA21125, NA21126, NA21127, NA21128, NA21129, NA21130, NA21133, NA21135, NA21137, NA21141, NA21143.

PJL includes 85 samples:

HG01583, HG01586, HG01589, HG01593, HG02490, HG02491, HG02493, HG02494, HG02597, HG02600, HG02601, HG02603, HG02648, HG02649, HG02651, HG02654, HG02655, HG02657, HG02658, HG02660, HG02661, HG02681, HG02682, HG02684, HG02685, HG02687, HG02688, HG02690, HG02694, HG02696, HG02697, HG02699,

HG02700, HG02724, HG02725, HG02727, HG02728, HG02731, HG02733, HG02734, HG02737, HG02778, HG02780, HG02783, HG02784, HG02786, HG02787, HG02789, HG02790, HG02792, HG02793, HG03016, HG03018, HG03019, HG03021, HG03022, HG03229, HG03234, HG03237, HG03238, HG03491, HG03619, HG03624, HG03625, HG03629, HG03631, HG03634, HG03636, HG03640, HG03649, HG03652, HG03653, HG03660, HG03663, HG03667, HG03668, HG03702, HG03703, HG03705, HG03706, HG03708, HG03709, HG03762, HG03765, HG03767.

BEB includes 68 samples:

HG03006, HG03007, HG03585, HG03589, HG03594, HG03595, HG03600, HG03604, HG03607, HG03615, HG03616, HG03793, HG03796, HG03800, HG03802, HG03803, HG03805, HG03808, HG03812, HG03814, HG03815, HG03821, HG03823, HG03824, HG03826, HG03829, HG03830, HG03832, HG03902, HG03905, HG03907, HG03908, HG03910, HG03913, HG03914, HG03917, HG03919, HG03922, HG03925, HG03926, HG03931, HG03937, HG03940, HG03941, HG04131, HG04134, HG04140, HG04141, HG04146, HG04152, HG04153, HG04155, HG04158, HG04161, HG04162, HG04164, HG04171, HG04173, HG04176, HG04177, HG04180, HG04182, HG04183, HG04185, HG04186, HG04188, HG04189, HG04195.

STU includes 91 samples:

HG03642, HG03643, HG03644, HG03646, HG03672, HG03673, HG03679, HG03680, HG03681, HG03684, HG03685, HG03686, HG03687, HG03689, HG03690, HG03691, HG03692, HG03693, HG03695, HG03697,HG03698, HG03733, HG03736, HG03740, HG03741, HG03743, HG03745, HG03746, HG03750, HG03752, HG03753, HG03754, HG03755, HG03756, HG03757, HG03760, HG03836, HG03837, HG03838, HG03844, HG03848, HG03849, HG03850, HG03851, HG03854, HG03856, HG03857, HG03858, HG03884, HG03885, HG03886, HG03887, HG03888, HG03890, HG03894, HG03895, HG03896, HG03897, HG03898, HG03900, HG03943, HG03944, HG03947, HG03948, HG03949, HG03950, HG03953, HG03955, HG03985, HG03986, HG03989, HG03990, HG03991, HG03995, HG03998, HG03999, HG04003, HG04006, HG04029, HG04033, HG04035, HG04038, HG04039, HG04042, HG04047, HG04075, HG04099, HG04100, HG04107, HG04227, HG04229.

ITU includes 86 samples:

HG03713, HG03714, HG03715, HG03716, HG03717, HG03722, HG03727, HG03729, HG03730, HG03731, HG03770, HG03771, HG03772, HG03773, HG03774, HG03775, HG03778, HG03779, HG03781, HG03784, HG03785, HG03786, HG03787, HG03788, HG03789, HG03790, HG03861, HG03862, HG03863, HG03864, HG03866, HG03867, HG03869, HG03870, HG03871, HG03872, HG03873, HG03874, HG03875, HG03882, HG03960, HG03963, HG03965, HG03967, HG03968, HG03969, HG03971, HG03973, HG03974, HG03976, HG03977, HG03978, HG04001, HG04002, HG04014, HG04015, HG04018, HG04019, HG04020, HG04022, HG04023, HG04025, HG04054, HG04056,

HG04059, HG04060, HG04061, HG04062, HG04063, HG04070, HG04076, HG04080, HG04090, HG04093, HG04094, HG04096, HG04098, HG04118, HG04200, HG04202, HG04206, HG04212, HG04219, HG04225, HG04235, HG04238.

The list of 451 samples from SRA classified by cancer types

Esophageal cancer includes 145 samples:

SRR1056623, SRR1056628, SRR1056629, SRR1056630, SRR1056631, SRR1056632, SRR1056633, SRR1056634, SRR1056635, SRR1056636, SRR1056637, SRR1056638, SRR1056639, SRR1056640, SRR1056641, SRR1056642, SRR1056643, SRR1056644, SRR1056645, SRR1056646, SRR1056647, SRR1056648, SRR1056649, SRR1056650, SRR1056651, SRR1056652, SRR1056653, SRR1056654, SRR1056655, SRR1056656, SRR1056657, SRR1056658, SRR1056659, SRR1056660, SRR1056661, SRR1056662, SRR1056663, SRR1056664, SRR1056665, SRR1056666, SRR1056667, SRR1056668, SRR1056669, SRR1056670, SRR1056671, SRR1056672, SRR1056673, SRR1056674, SRR1056675, SRR1056676, SRR1056677, SRR1056679, SRR1056680, SRR1056681, SRR1056682, SRR1056683, SRR1056684, SRR1056685, SRR1056686, SRR1056687, SRR1056688, SRR1056689, SRR1056690, SRR1056691, SRR1056692, SRR1056693, SRR1056694, SRR1056695, SRR1056696, SRR1056697, SRR1056698, SRR1056699, SRR1056703, SRR1056705, SRR1056706, SRR1056708, SRR1056710, SRR1056711, SRR1056712, SRR1056713, SRR1056715, SRR1056721, SRR1056722, SRR1056723, SRR1056724, SRR1056725, SRR1056726, SRR1056727, SRR1056728, SRR1056729,

SRR1056730, SRR1056731, SRR1056732, SRR1056733, SRR1056734, SRR1056735, SRR1056736, SRR1056737, SRR1056738, SRR1056739, SRR1056740, SRR1056741, SRR1056742, SRR1056743, SRR1056744, SRR1056745, SRR1056747, SRR1056748, SRR1056749, SRR1056750, SRR1056751, SRR1056752, SRR1056753, SRR1056754, SRR1056755, SRR1056756, SRR1056757, SRR1056758, SRR1056759, SRR1056760, SRR1056761, SRR1056762, SRR1056763, SRR1056764, SRR1056765, SRR1056766, SRR1056767, SRR1056768, SRR1056769, SRR1056770, SRR1056771, SRR1056772, SRR1056773, SRR1056774, SRR1056775, SRR1056776, SRR1056777, SRR1056778, SRR1056779, SRR1056780, SRR1056781, SRR1056782, SRR1056783, SRR1056784, SRR1056785.

Bladder cancer includes 131 samples:

SRR645164, SRR645165, SRR645166, SRR645167, SRR645168, SRR645169, SRR645170, SRR645171, SRR645172, SRR645173, SRR645174, SRR645175, SRR645176, SRR645177, SRR645178, SRR645179, SRR645180, SRR645181, SRR645182, SRR645183, SRR645184, SRR645185, SRR645186, SRR645187, SRR645188, SRR645189, SRR645190, SRR645191, SRR645192, SRR645193, SRR645194, SRR645201, SRR645203, SRR645205, SRR645207, SRR645209, SRR645213, SRR645214, SRR645216, SRR645218, SRR645220, SRR645222, SRR645224, SRR645226, SRR645227, SRR645229, SRR645231, SRR645233, SRR645235, SRR645237, SRR645239, SRR645241, SRR645243, SRR645245, SRR645247, SRR645249, SRR645253, SRR645255, SRR645257, SRR645259,

SRR645264, SRR645267, SRR645269, SRR645270, SRR645272, SRR645274, SRR645276, SRR645278, SRR645281, SRR645283, SRR645284, SRR645285, SRR645286, SRR645287, SRR645288, SRR645289, SRR645290, SRR645291, SRR645292, SRR645293, SRR645294, SRR645295, SRR645296, SRR645297, SRR645298, SRR645299, SRR645300, SRR645302, SRR645305, SRR645306, SRR645307, SRR645309, SRR645310, SRR645311, SRR645312, SRR645313, SRR645314, SRR645315, SRR645317, SRR645318, SRR645320, SRR645322, SRR645323, SRR645324, SRR645326, SRR645328, SRR645330, SRR645332, SRR645335, SRR645338, SRR645341, SRR645343, SRR645345, SRR645347, SRR645348, SRR645349, SRR645351, SRR645352, SRR645353, SRR645355, SRR645357, SRR645359, SRR645362, SRR645364, SRR645365, SRR645366, SRR645368, SRR645370, SRR645372, SRR645373, SRR645374.

Hepatocellular carcinoma includes 67 samples:

ERR232258, ERR232259, ERR232260, ERR232261, ERR232262, ERR232263, ERR232264, ERR232265, ERR232266, ERR232267, ERR232268, ERR232269, ERR232270, ERR232271, ERR232272, ERR232273, ERR232274, ERR232275, ERR232276, ERR232277, ERR232278, ERR232279, SRR2049021, SRR2049022, SRR2157816, SRR2157817, SRR2157818, SRR2157821, SRR2157822, SRR2157823, SRR2157824, SRR2157825, SRR2157826, SRR2157827, SRR2157828, SRR2157829, SRR2157831, SRR2157832, SRR2157833, SRR2157834, SRR2157835, SRR2157836, SRR2157837, SRR2157838, SRR2157839, SRR2157842, SRR2157843, SRR2157844,

SRR2157845, SRR2157846, SRR2157848, SRR2157849, SRR2157850, SRR2157851, SRR2157853, SRR2157854, SRR2157855, SRR2157856, SRR2157857, SRR2157859, SRR2157860, SRR2157861, SRR2157862, SRR2157863, SRR2157864, SRR2157867, SRR5296488.

Lung cancer includes 62 samples:

SRR975195, SRR975201, SRR975209, SRR975213, SRR975215, SRR975217, SRR975219, SRR975221, SRR975225, SRR975229, SRR975233, SRR975235, SRR975237, SRR975239, SRR975241, SRR975243, SRR975245, SRR975249, SRR975251, SRR975253, SRR975255, SRR975257, SRR975261, SRR975262, SRR975263, SRR975264, SRR975265, SRR975272, SRR975274, SRR975276, SRR975282, SRR975284, SRR975286, SRR975288, SRR975290, SRR975292, SRR975294, SRR975296, SRR975304, SRR975306, SRR975308, SRR975310, SRR975318, SRR975320, SRR975322, SRR975323, SRR975324, SRR975325, SRR975327, SRR975328, SRR975329, SRR975330, SRR975331, SRR975333, SRR975334, SRR975335, SRR975336, SRR975337, SRR975338, SRR975339, SRR975341, SRR981030.

Prostate cancer includes 32 samples:

ERR532450, ERR532451, ERR532458, ERR532459, ERR532460, ERR532461, ERR532462, ERR532465, ERR532503, ERR532504, ERR532505, ERR532522, ERR532523, ERR532524, ERR532525, ERR532526, ERR532527, ERR532528, ERR532529, ERR532530, ERR532531, ERR532532, ERR532533, ERR532534, ERR532535, ERR532536, ERR532537, ERR532538, ERR532539, ERR532540, ERR532541, ERR532542.

Pancreatic cancer includes 14 samples:

ERR232239, ERR232240, ERR232241, ERR232242, ERR232243, ERR232244, ERR232245, ERR232246, ERR232247, ERR232248, ERR232249, ERR232250, ERR232251, ERR232252.

APPENDIX B. PUBLICATIONS

[1] **Qu L**, Wang L, He F, Han Y, Yang L, & Zhu H*. Landscape of Micro-inversions with Clue to Population Genetics Analysis in Human Genomes. (Submitted to American Journal of Human Genetics).

[2] **Qu L**, Zhu H, & Wang M D*. Micro-Inversions in Human Cancer Genomes. The 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, 2018: 1323-1326. (EI Index)

[3] **Qu L**, Wang L, He F, Han Y, Yang L, & Zhu H*. The analysis of micro-inversions in human genomes. The 13th annual meeting of the Chinese society of bioengineering and the 2019 national biotechnology conference. 2019.

[4] Wang L, **Qu L**, Yang L, Wang Y, & Zhu H*. NanoReviser: An error-correction tool for nanopore sequencing based on deep learning algorithm. (Submitted to Frontiers in Genetics).

[5] Hoffman R A, Venugopalan J, **Qu L**, Wu H, &Wang M D*. Improving Validity of Cause of Death on Death Certificates. Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics. ACM, 2018: 178-183. (EI Index)

REFERENCES

[1] Chaisson M J P, Sanders A D, Zhao X, et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. Nature Communications, 2019, 10.

[2] Sequencing T C, Waterson R H, Lander E S, et al. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature, 2005, 437(7055): 69.

[3] Gibbs R A, Rogers J, Katze M G, et al. Evolutionary and biomedical insights from the rhesus macaque genome. Science, 2007, 316(5822): 222-234.

[4] Han K, Lee J, Meyer T J, et al. Alu recombination-mediated structural deletions in the chimpanzee genome. PLoS Genetics, 2007, 3(10): e184.

[5] Kehrer-Sawatzki H, Cooper D N. Molecular mechanisms of chromosomal rearrangement during primate evolution. Chromosome Research, 2008, 16(1): 41-56.

[6] Lowry D B, Willis J H. A widespread chromosomal inversion polymorphism contributes to a major life-history transition, local adaptation, and reproductive isolation. PLoS Biology, 2010, 8(9): e1000500.

[7] Baker M. Structural variation: the genome's hidden architecture. Nature Methods, 2012, 9(2): 133-137.

[8] Collins R L, Brand H, Redin C E, et al. Defining the diverse spectrum of inversions, complex structural variation, and chromothripsis in the morbid human genome. Genome Biology, 2017, 18(1): 36.

[9] Pang A W, MacDonald J R, Pinto D, et al. Towards a comprehensive structural variation map of an individual human genome. Genome Biology, 2010, 11(5): R52.

[10] Weckselblatt B, Rudd M K. Human structural variation: mechanisms of chromosome rearrangements. Trends in Genetics, 2015, 31(10): 587-599.

[11] Karayiorgou M, Simon T J, Gogos J A. 22q11 2 microdeletions: linking DNA structural variation to brain dysfunction and schizophrenia. Nature Reviews Neuroscience, 2010, 11(6): 402.

[12] Werling D M, Brand H, An J Y, et al. An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder. Nature Genetics, 2018, 50(5): 727.

[13] Quigley D A, Dang H X, Zhao S G, et al. Genomic hallmarks and structural variation in metastatic prostate cancer. Cell, 2018, 174(3): 758-769 e9.

[14] Saadati H R, Wittig M, Helbig I, et al. Genome-wide rare copy number variation screening in ulcerative colitis identifies potential susceptibility loci. BMC Medical Genetics, 2016, 17(1): 26.

[15] Lupski J R. Structural variation mutagenesis of the human genome: Impact on disease and evolution. Environmental and Molecular Mutagenesis, 2015, 56(5): 419-436.

[16] Medvedev P, Stanciu M, Brudno M. Computational methods for discovering structural variation with next-generation sequencing. Nature Methods, 2009, 6(11s): S13.

[17] Goodwin S, McPherson J D, McCombie W R. Coming of age: ten years of next-generation sequencing technologies. Nature Reviews Genetics, 2016, 17(6): 333.

[18] Green E D, Watson J D, Collins F S. Human Genome Project: Twenty-five years of big biology. Nature News, 2015, 526(7571): 29.

[19] Shumway M, Cochrane G, Sugawara H. Archiving next generation sequencing data. Nucleic Acids Research, 2009, 38(suppl_1): D870-D871.

[20] Sharp A J, Cheng Z, Eichler E E. Structural variation of the human genome. Annual Review of Genomics and Human Genetics, 2006, 7: 407-442.

[21] Tan O, Shrestha R, Cunich M, et al. Application of next-generation sequencing to improve cancer management: A review of the clinical effectiveness and cost-effectiveness. Clinical Genetics, 2018, 93(3): 533-544.

[22] Jones M, Zheng Z, Wang J, et al. Impact of next-generation sequencing on the clinical diagnosis of pancreatic cysts. Gastrointestinal Endoscopy, 2016, 83(1): 140-148.

[23] Girirajan S, Eichler E E. Phenotypic variability and genetic susceptibility to genomic disorders. Human Molecular Genetics, 2010, 19(R2): R176-R187.

[24] Korbel J O, Urban A E, Affourtit J P, et al. Paired-end mapping reveals extensive structural variation in the human genome. Science, 2007, 318(5849): 420-426.

[25] Sebat J, Lakshmi B, Troge J, et al. Large-scale copy number polymorphism in the human genome. Science, 2004, 305(5683): 525-528.

[26] Zhang J, Chiodini R, Badr A, et al. The impact of next-generation sequencing on genomics. Journal of Genetics and Genomics, 2011, 38(3): 95-109.

[27] Zhao M, Wang Q, Wang Q, et al. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. BMC Bioinformatics, 2013, 14(11): S1.

[28] Brandler W M, Antaki D, Gujral M, et al. Paternally inherited cis-regulatory structural variants are associated with autism. Science, 2018, 360(6386): 327-331.

[29] van der Merwe C, Carr J, Glanzmann B, et al. Exonic rearrangements in the known Parkinson's disease-causing genes are a rare cause of the disease in South African patients. Neuroscience Letters, 2016, 619: 168-171.

[30] Butcher N J, Merico D, Zarrei M, et al. Whole-genome sequencing suggests mechanisms for 22q11 2 deletion-associated Parkinson's disease. PLoS ONE, 2017, 12(4): e0173944.

[31] Park C Y, Kim D H, Son J S, et al. Functional correction of large factor VIII gene chromosomal inversions in hemophilia A patient-derived iPSCs using CRISPR-Cas9. Cell Stem Cell, 2015, 17(2): 213-220.

[32] Stankiewicz P, Lupski J R. Structural variation in the human genome and its role in disease. Annual Review of Medicine, 2010, 61: 437-455.

[33] Duraisingh M T, Triglia T, Cowman A F. Negative selection of Plasmodium falciparum reveals targeted gene deletion by double crossover recombination. International Journal for Parasitology, 2002, 32(1): 81-89.

[34] Luo S, Jane A Y, Song Y S. Estimating copy number and allelic variation at the immunoglobulin heavy chain locus using short reads. PLoS Computational Biology, 2016, 12(9): e1005117.

[35] Tuzun E, Sharp A J, Bailey J A, et al. Fine-scale structural variation of the human genome. Nature Genetics, 2005, 37(7): 727.

[36] Pagel K A, Antaki D, Lian A J, et al. Pathogenicity and functional impact of non-frameshifting insertion/deletion variation in the human genome. PLoS Computational Biology, 2019, 15(6): e1007112.

[37] Onozawa M, Goldberg L, Aplan P D. Landscape of insertion polymorphisms in the human genome. Genome Biology and Evolution, 2015, 7(4): 960-968.

[38] Wang Y, Su P, Hu B, et al. Characterization of 26 deletion CNVs reveals the frequent occurrence of micro-mutations within the breakpoint-flanking regions and frequent repair of double-strand breaks by templated insertions derived from remote genomic regions. Human Genetics, 2015, 134(6): 589-603.

[39] Vaszkó T, Papp J, Krausz C, et al. Discrimination of deletion and duplication subtypes of the deleted in azoospermia gene family in the context of frequent interloci gene conversion. PLoS ONE, 2016, 11(10): e0163936.

[40] Malekpour S A, Pezeshk H, Sadeghi M. PSE-HMM: genome-wide CNV detection from NGS data using an HMM with Position-Specific Emission probabilities. BMC Bioinformatics, 2017, 18(1): 30.

[41] Zhou B, Ho S S, Zhang X, et al. Whole-genome sequencing analysis of CNV using

low-coverage and paired-end strategies is efficient and outperforms array-based CNV analysis. Journal of Medical Genetics, 2018, 55(11): 735-743.

[42] Conrad D F, Hurles M E. The population genetics of structural variation. Nature Genetics, 2007, 39(7s): S30.

[43] He F, Li Y, Tang Y H, et al. Identifying micro-inversions using high-throughput sequencing reads. BMC Genomics, 2016, 17(1): 4.

[44] Catacchio C R, Maggiolini F A M, D'Addabbo P, et al. Inversion variants in human and primate genomes. Genome Research, 2018, 28(6): 910-920.

[45] Feuk L, MacDonald J R, Tang T, et al. Discovery of human inversion polymorphisms by comparative analysis of human and chimpanzee DNA sequence assemblies. PLoS Genetics, 2005, 1(4): e56..

[46] Bansal V, Bashir A, Bafna V. Evidence for large inversion polymorphisms in the human genome from HapMap data. Genome Research, 2007, 17(2): 219-230.

[47] Pittman A M, Myers A J, Duckworth J, et al. The structure of the tau haplotype in controls and in progressive supranuclear palsy. Human Molecular Genetics, 2004, 13(12): 1267-1274.

[48] Skipper L, Wilkes K, Toft M, et al. Linkage disequilibrium and association of MAPT H1 in Parkinson disease. The American Journal of Human Genetics, 2004, 75(4): 669-677.

[49] Flores M, Morales L, Gonzaga-Jauregui C, et al. Recurrent DNA inversion rearrangements in the human genome. Proceedings of the National Academy of Sciences, 2007, 104(15): 6099-6106.

[50] Pang A W C, Migita O, MacDonald J R, et al. Mechanisms of formation of structural variation in a fully sequenced human genome. Human Mutation, 2013, 34(2): 345-354.

[51] Escaram ś G, Docampo E, Rabionet R. A decade of structural variants: description, history and methods to detect structural variation. Briefings in Functional Genomics, 2015, 14(5): 305-314.

[52] Puig M, Casillas S, Villatoro S, et al. Human inversions and their functional consequences. Briefings in Functional Genomics, 2015, 14(5): 369-379.

[53] C áceres A, Sindi S S, Raphael B J, et al. Identification of polymorphic inversions from genotypes. BMC Bioinformatics, 2012, 13(1): 28.

[54] Pehlivan T, Pober B R, Brueckner M, et al. GATA4 haploinsufficiency in patients with interstitial deletion of chromosome region 8p23.1 and congenital heart disease. American Journal of Medical Genetics, 1999, 83(3): 201-206.

[55] Salm M P A, Horswell S D, Hutchison C E, et al. The origin, global distribution, and functional impact of the human 8p23 inversion polymorphism. Genome Research, 2012, 22(6): 1144-1153.

[56] Navarro A, Barton N H. Chromosomal speciation and molecular divergence--accelerated evolution in rearranged chromosomes. Science, 2003, 300(5617): 321-324.

[57] Osborne L R, Li M, Pober B, et al. A 1.5 million–base pair inversion polymorphism in families with Williams-Beuren syndrome. Nature Genetics, 2001, 29(3): 321.

[58] Thomas N S, Bryant V, Maloney V, et al. Investigation of the origins of human autosomal inversions. Human Genetics, 2008, 123(6): 607-616.

[59] Giglio S, Calvari V, Gregato G, et al. Heterozygous submicroscopic inversions involving olfactory receptor–gene clusters mediate the recurrent t (4; 8)(p16; p23) translocation. The American Journal of Human Genetics, 2002, 71(2): 276-285.

[60] Antonacci F, Kidd J M, Marques-Bonet T, et al. Characterization of six human disease-associated inversion polymorphisms. Human Molecular Genetics, 2009, 18(14): 2555-2566.

[61] Small K, Iber J, Warren S T. Emerin deletion reveals a common X-chromosome inversion mediated by inverted repeats. Nature Genetics, 1997, 16(1): 96.

[62] Jaeken J, Martens K, Fran ̧ois I, et al. Deletion of PREPL, a gene encoding a putative serine oligopeptidase, in patients with hypotonia-cystinuria syndrome. The American Journal of Human Genetics, 2006, 78(1): 38-51.

[63] Levy S, Sutton G, Ng P C, et al. The diploid genome sequence of an individual human. PLoS Biology, 2007, 5(10): e254.

[64] Alves J M, Chikhi L, Amorim A, et al. The 8p23 inversion polymorphism determines local recombination heterogeneity across human populations. Genome Biology and Evolution, 2014, 6(4): 921-930.

[65] Muthuvel A, Ravindran M, Chander A, et al. Pericentric inversion of chromosome 9 causing infertility and subsequent successful in vitro fertilization. Nigerian Medical Journal: Journal of the Nigeria Medical Association, 2016, 57(2): 142.

[66] Feuk L. Inversion variants in the human genome: role in disease and genome architecture. Genome Medicine, 2010, 2(2): 11.

[67] Feuk L, Carson A R, Scherer S W. Structural variation in the human genome. Nature Reviews Genetics, 2006, 7(2): 85.

[68] Church D M, Lappalainen I, Sneddon T P, et al. Public data archives for genomic structural variation. Nature Genetics, 2010, 42(10): 813.

[69] MacDonald J R, Ziman R, Yuen R K C, et al. The Database of Genomic Variants: a curated collection of structural variation in the human genome. Nucleic Acids Research, 2013, 42(D1): D986-D992.

[70] Conrad D F, Andrews T D, Carter N P, et al. A high-resolution survey of deletion polymorphism in the human genome. Nature Genetics, 2006, 38(1): 75.

[71] Hurles M E, Dermitzakis E T, Tyler-Smith C. The functional impact of structural variation in humans. Trends in Genetics, 2008, 24(5): 238-245.

[72] Chen R, Lau Y L, Zhang Y , et al. SRinversion: a tool for detecting short inversions by splitting and re-aligning poorly mapped and unmapped sequencing reads. Bioinformatics, 2016, 32(23): 3559-3565.

[73] Tattini L, D'Aurizio R, Magi A. Detection of genomic structural variants from next-generation sequencing data. Frontiers in Bioengineering and Biotechnology, 2015, 3: 92.

[74] Sanger F, Coulson A R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. Journal of Molecular Biology, 1975, 94(3): 441-448.

[75] DePristo M A, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature Genetics, 2011, 43(5): 491.

[76] Venter J C, Smith H O, Hood L. A new strategy for genome sequencing. Nature, 1996, 381(6581): 364.

[77] Sims D, Sudbery I, Ilott N E, et al. Sequencing depth and coverage: key considerations in genomic analyses. Nature Reviews Genetics, 2014, 15(2): 121.

[78] Metzker M L. Sequencing technologies—the next generation. Nature Reviews Genetics, 2010, 11(1): 31.

[79] Mardis E R. The impact of next-generation sequencing technology on genetics. Trends in Genetics, 2008, 24(3): 133-141.

[80] Ng P C, Kirkness E F. Whole genome sequencing. Humana Press, Totowa, NJ, 2010: 215-226.

[81] Eck S H, Benet-Pagès A, Flisikowski K, et al. Whole genome sequencing of a single Bos taurus animal for single nucleotide polymorphism discovery. Genome Biology, 2009, 10(8): R82.

[82] Cao J, Schneeberger K, Ossowski S, et al. Whole-genome sequencing of multiple Arabidopsis thaliana populations. Nature Genetics, 2011, 43(10): 956.

[83] Alkan C, Ventura M, Archidiacono N, et al. Organization and evolution of primate centromeric DNA from whole-genome shotgun sequence data. PLoS Computational Biology, 2007, 3(9): e181.

[84] Gilissen C, Hehir-Kwa J Y, Thung D T, et al. Genome sequencing identifies major causes of severe intellectual disability. Nature, 2014, 511(7509): 344.

[85] Van El C G, Cornel M C, Borry P, et al. Whole-genome sequencing in health care. European Journal of Human Genetics, 2013, 21(6): 580.

[86] Huang X, Feng Q, Qian Q, et al. High-throughput genotyping by whole-genome resequencing. Genome Research, 2009, 19(6): 1068-1076.

[87] Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics, 2009, 25(14): 1754-1760.

[88] Abyzov A, Urban A E, Snyder M, et al. CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. Genome Research, 2011, 21(6): 974-984.

[89] Rausch T, Zichner T, Schlattl A, et al. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics, 2012, 28(18): i333-i339.

[90] Ye K, Schulz M H, Long Q, et al. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics, 2009, 25(21): 2865-2871.

[91] Lin K, Smit S, Bonnema G, et al. Making the difference: integrating structural variation detection tools. Briefings in Bioinformatics, 2014, 16(5): 852-864.

[92] Hara Y, Imanishi T. Abundance of ultramicro inversions within local alignments between human and chimpanzee genomes. BMC Evolutionary Biology, 2011, 11(1): 308.

[93] Yunis J J, Sawyer J R, Dunham K. The striking resemblance of high-resolution G-banded chromosomes of man and chimpanzee. Science, 1980: 1145-1148..

[94] McLysaght A, Seoighe C, Wolfe K H. High frequency of inversions during eukaryote gene order evolution. Springer Dordrecht, 2000: 47-58.

[95] Stefansson H, Helgason A, Thorleifsson G, et al. A common inversion under selection in Europeans. Nature Genetics, 2005, 37(2): 129.

[96] Kidd J M, Cooper G M, Donahue W F, et al. Mapping and sequencing of structural variation from eight human genomes. Nature, 2008, 453(7191): 56.

[97] Baker M, Litvan I, Houlden H, et al. Association of an extended haplotype in the tau

gene with progressive supranuclear palsy. Human Molecular Genetics, 1999, 8(4): 711-715.

[98] Oliveira S A, Scott W K, Zhang F, et al. Linkage disequilibrium and haplotype tagging polymorphisms in the Tau H1 haplotype. Neurogenetics, 2004, 5(3): 147-155.

[99] Jobling M A, Williams G, Schiebel K, et al. A selective difference between human Y-chromosomal DNA haplotypes. Current Biology, 1998, 8(25): 1391-1394.

[100] Kurotaki N, Imaizumi K, Harada N, et al. Haploinsufficiency of NSD1 causes Sotos syndrome. Nature Genetics, 2002, 30(4): 365.

[101] Kolb J, Chuzhanova N A, Högel J, et al. Cruciform-forming inverted repeats appear to have mediated many of the microinversions that distinguish the human and chimpanzee genomes. Chromosome Research, 2009, 17(4): 469-483.

[102] Hou M, Yao P, Antonou A, et al. Pico-inplace-inversions between human and chimpanzee. Bioinformatics, 2011, 27(23): 3266-3275.

[103] Chaisson M J, Raphael B J, Pevzner P A. Microinversions in mammalian evolution. Proceedings of the National Academy of Sciences, 2006, 103(52): 19824-19829.

[104] Fischer G, Rocha E P C, Brunet F, et al. Highly variable rates of genome rearrangements between hemiascomycetous yeast lineages. PLoS Genetics, 2006, 2(3): e32.

[105] Jones F C, Grabherr M G, Chan Y F, et al. The genomic basis of adaptive evolution in threespine sticklebacks. Nature, 2012, 484(7392): 55.

[106] Groeters F R, Shaw D D. Association between latitudinal variation for embryonic development time and chromosome structure in the grasshopper Caledia captiva (Orthoptera: Acrididae). Evolution, 1992, 46(1): 245-257.

[107] Kennington W J, Partridge L, Hoffmann A A. Patterns of diversity and linkage disequilibrium within the cosmopolitan inversion In (3R) Payne in Drosophila melanogaster are indicative of coadaptation. Genetics, 2006, 172(3): 1655-1663.

[108] Hoffmann A A, Rieseberg L H. Revisiting the impact of inversions in evolution: from population genetic markers to drivers of adaptive shifts and speciation? Annual Review of Ecology, Evolution, and Systematics, 2008, 39: 21-42.

[109] Bourque G, Zdobnov E M, Bork P, et al. Comparative architectures of mammalian and chicken genomes reveal highly variable rates of genomic rearrangements across different lineages. Genome Research, 2005, 15(1): 98-110.

[110] Braun E L, Kimball R T, Han K L, et al. Homoplastic microinversions and the avian tree of life. BMC Evolutionary Biology, 2011, 11(1): 141.

[111] Lyon M F. Transmission ratio distortion in mice. Annual Review of Genetics, 2003, 37(1): 393-408.

[112] M Alves J, M Lopes A, Chikhi L, et al. On the structural plasticity of the human genome: chromosomal inversions revisited. Current Genomics, 2012, 13(8): 623-632.

[113] Stringer C B, Andrews P. Genetic and fossil evidence for the origin of modern humans. Science, 1988, 239(4845): 1263-1268.

[114] 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. Nature, 2012, 491(7422): 56.

[115] Sudmant P H, Rausch T, Gardner E J, et al. An integrated map of structural variation in 2,504 human genomes. Nature, 2015, 526(7571): 75.

[116] 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature, 2015, 526(7571): 68.

[117] Pemberton T J, Szpiech Z A. Relationship between deleterious variation, genomic autozygosity, and disease risk: insights from The 1000 Genomes Project. The American Journal of Human Genetics, 2018, 102(4): 658-675.

[118] Peng T, Wang L, Li G. The analysis of APOL1 genetic variation and haplotype diversity provided by 1000 Genomes project. BMC Nephrology, 2017, 18(1): 267.

[119] Stringer C. Palaeoanthropology: coasting out of Africa. Nature, 2000, 405(6782): 24.

[120] Karolchik D, Baertsch R, Diekhans M, et al. The UCSC genome browser database. Nucleic Acids Research, 2003, 31(1): 51-54.

[121] Harrow J, Frankish A, Gonzalez J M, et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Research, 2012, 22(9): 1760-1774.

[122] Safran M, Dalah I, Alexander J, et al. GeneCards Version 3: the human gene integrator. Database, 2010, 2010.

[123] Zhou Y, Zhou B, Pache L, et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nature Communications, 2019, 10(1): 1523.

[124] Lou H, Li S, Yang Y, et al. A map of copy number variations in Chinese populations. PLoS ONE, 2011, 6(11): e27341.

[125] Jo B S, Choi S S. Introns: the functional benefits of introns in genomes. Genomics & Informatics, 2015, 13(4): 112.

[126] Hughes T A. Regulation of gene expression by alternative untranslated regions. Trends in Genetics, 2006, 22(3): 119-122.

[127] Xu K, Kranzler H R, Sherva R, et al. Genomewide association study for maximum number of alcoholic drinks in European Americans and African Americans. Alcoholism: Clinical and Experimental Research, 2015, 39(7): 1137-1147.

[128] Yuan A, Yi Z, Wang Q, et al. ANK3 as a risk gene for schizophrenia: new data in Han Chinese and meta analysis. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics, 2012, 159(8): 997-1005.

[129] Harding R M, Fullerton S M, Griffiths R C, et al. Archaic African and Asian lineages in the genetic ancestry of modern humans. American Journal of Human Genetics, 1997, 60(4): 772.

[130] Baharian S, Barakatt M, Gignoux C R, et al. The great migration and African-American genomic diversity. PLoS Genetics, 2016, 12(5): e1006059.

[131] Via M, Gignoux C R, Roth L A, et al. History shaped the geographic distribution of genomic admixture on the island of Puerto Rico. PLoS One, 2011, 6(1): e16513.

[132] Cobo M F, Jobes D V, Yanagihara R, et al. Reconstructing population history using JC virus: Amerinds, Spanish, and Africans in the ancestry of modern Puerto Ricans. Human Biology, 2001, 73(3): 385-402.

[133] May A, Hazelhurst S, Li Y, et al. Genetic diversity in black South Africans from Soweto. BMC Genomics, 2013, 14(1): 644.

[134] Lao O, Lu T T, Nothnagel M, et al. Correlation between genetic and geographic structure in Europe. Current Biology, 2008, 18(16): 1241-1248.

[135] Takeuchi F, Katsuya T, Kimura R, et al. The fine-scale genetic structure and evolution of the Japanese population. PLoS ONE, 2017, 12(11): e0185487.

[136] Jeong C, Nakagome S, Di Rienzo. A Deep history of East Asian populations revealed through genetic analysis of the Ainu. Genetics, 2016, 202(1): 261-272.

[137] Council H A. Race & ethnicity in rural America. Rural Research Briefs, April, 2012: 1283-88.

[138] Fage J D. Slavery and the slave trade in the context of West African history. The Journal of African History, 1969, 10(3): 393-404.

[139] Djamba Y K. African immigrants in the United States: A socio-demographic profile in comparison to native blacks. Journal of Asian and African Studies, 1999, 34(2): 210-215.

[140] Zhang Y, Li S, Abyzov A, et al. Landscape and variation of novel retroduplications in 26 human populations. PLoS Computational Biology, 2017, 13(6): e1005567.

[141] Palo J U, Ulmanen I, Lukka M, et al. Genetic markers and population history: Finland revisited European. Journal of Human Genetics, 2009, 17(10): 1336.

[142] Takenaka A. The Japanese in Peru: History of immigration, settlement, and racialization Latin. American Perspectives, 2004, 31(3): 77-98.

[143] Templeton A. Out of Africa again and again. Nature, 2002, 416(6876): 45.

[144] Macaulay V, Hill C, Achilli A, et al. Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. Science, 2005, 308(5724): 1034-1036.

[145] Hollfelder N, Schlebusch C M, Günther T, et al. Northeast African genomic variation shaped by the continuity of indigenous groups and Eurasian migrations. PLoS Genetics, 2017, 13(8): e1006976.

[146] Kim B J, Kim A R, Han J H, et al. Discovery of MYH14 as an important and unique deafness gene causing prelingually severe autosomal dominant nonsyndromic hearing loss. The Journal of Gene Medicine, 2017, 19(4): e2950.

[147] Friedman R A, Van Laer L, Huentelman M J, et al. GRM7 variants confer susceptibility to age-related hearing impairment. Human Molecular Genetics, 2008, 18(4): 785-796.

[148] De Beeck K O, Van Camp G, Thys S, et al. The DFNA5 gene, responsible for hearing loss and involved in cancer, encodes a novel apoptosis-inducing protein. European Journal of Human Genetics, 2011, 19(9): 965.

[149] Church D M, Schneider V A, Graves T, et al. Modernizing reference genome assemblies. PLoS Biology, 2011, 9(7): e1001091.

[150] Stratton M R, Campbell P J, Futreal P A. The cancer genome. Nature, 2009, 458(7239): 719.

[151] Czubak K, Lewandowska M A, Klonowska K, et al. High copy number variation of cancer-related microRNA genes and frequent amplification of DICER1 and DROSHA in lung cancer. Oncotarget, 2015, 6(27): 23399.

[152] Liu T C, Vachharajani N, Chapman W C, et al. SALL4 Immunoreactivity Predicts Prognosis in Western Hepatocellular Carcinoma Patients but is a Rare Event-A Study of 236 Cases. The American Journal of Surgical Pathology, 2014, 38(7): 966.

[153] Heestand G M, Kurzrock R. Molecular landscape of pancreatic cancer: implications for current clinical trials. Oncotarget, 2015, 6(7): 4553.

[154] Papaemmanuil E, Rapado I, Li Y, et al. RAG-mediated recombination is the predominant driver of oncogenic rearrangement in ETV6-RUNX1 acute

lymphoblastic leukemia. Nature Genetics, 2014, 46(2): 116.

[155] Stratton M R. Exploring the genomes of cancer cells: progress and promise. Science, 2011, 331(6024): 1553-1558.

[156] Abe S, Miura K, Kinoshita A, et al. Copy number variation of the antimicrobial-gene, defensin beta 4, is associated with susceptibility to cervical cancer. Journal of Human Genetics, 2013, 58(5): 250.

[157] Chalmers Z R, Connelly C F, Fabrizio D, et al. Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. Genome Medicine, 2017, 9(1): 34.

[158] Tubio J M C. Somatic structural variation and cancer. Briefings in Functional Genomics, 2015, 14(5): 339-351.

[159] Mardis E R, Wilson R K. Cancer genome sequencing: a review. Human Molecular Genetics, 2009, 18(R2): R163-R168.

[160] Barretina J, Caponigro G, Stransky N, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature, 2012, 483(7391): 603.

[161] Garralda E, Paz K, López-Casas P P, et al. Integrated next-generation sequencing and avatar mouse models for personalized cancer treatment. Clinical Cancer Research, 2014, 20(9): 2476-2484.

[162] Forbes S A, Beare D, Gunasekaran P, et al. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Research, 2014, 43(D1): D805-D811.

[163] Zhao H, Sun Z, Wang J, et al. CrossMap: a versatile tool for coordinate conversion between genome assemblies. Bioinformatics, 2013, 30(7): 1006-1007.

[164] Parris T Z, Aziz L, Kovács A, et al. Clinical relevance of breast cancer-related genes as potential biomarkers for oral squamous cell carcinoma. BMC Cancer, 2014, 14(1): 324.

[165] Smith D I, Zhu Y, McAvoy S, et al. Common fragile sites, extremely large genes, neural development and cancer. Cancer Letters, 2006, 232(1): 48-57.

[166] Masuda M, Uno Y, Ohbayashi N, et al. TNIK inhibition abrogates colorectal cancer stemness. Nature Communications, 2016, 7: 12586.

[167] Cao Q H, Liu F, Yang Z L, et al. Prognostic value of autophagy related proteins ULK1, Beclin 1, ATG3, ATG5, ATG7, ATG9, ATG10, ATG12, LC3B and p62/SQSTM1 in gastric cancer. American Journal of Translational Research, 2016, 8(9): 3831.

[168] Coolen M W, Stirzaker C, Song J Z, et al. Consolidation of the cancer genome into domains of repressive chromatin by long-range epigenetic silencing (LRES) reduces transcriptional plasticity. Nature Cell Biology, 2010, 12(3): 235.

[169] Poli G, Brancorsini S, Cochetti G, et al. Expression of inflammasome-related genes in bladder cancer and their association with cytokeratin 20 messenger RNA. Urologic Oncology: Seminars and Original Investigations Elsevier, 2015, 33(12): 505 e1-505 e7.

[170] Yen W H, Ke W S, Hung J J, et al. Sp1 - mediated ectopic expression of T-cell lymphoma invasion and metastasis 2 in hepatocellular carcinoma. Cancer Medicine, 2016, 5(3): 465-477.

[171] Li H, Zhao X, Li C, et al. Integrated analysis of lncRNA-associated ceRNA network reveals potential biomarkers for the prognosis of hepatitis B virus-related hepatocellular carcinoma. Cancer Management and Research, 2019, 11: 877.

[172] ESCUDERO-ESPARZA A, MARTIN T A, DAVIES M L, et al. PGF isoforms, PLGF-1 and PGF-2, in colorectal cancer and the prognostic significance. Cancer Genomics-Proteomics, 2009, 6(4): 239-246.

[173] Siggens L, Ekwall K. Epigenetics, chromatin and genome organization: recent advances from the ENCODE project. Journal of Internal Medicine, 2014, 276(3): 201-214.

[174] Scacheri C A, Scacheri P C. Mutations in the non-coding genome. Current Opinion in Pediatrics, 2015, 27(6): 659.